AQUATIC EFFECTS TECHNOLOGY EVALUATION (AETE) PROGRAM

Technical Evaluation on Methods for Benthic Invertebrate Data Analysis and Interpretation

AETE Project 2.1.3

104-03032/

TECHNICAL EVALUATION ON METHODS FOR BENTHIC INVERTEBRATE DATA ANALYSIS AND INTERPRETATION

FINAL REPORT

Barry R. Taylor
Taylor Mazier Associates
R.R. # 3, St. Andrews, N.S.

and

Robert C. Bailey,
Department of Zoology,
University of Western Ontario
London, Ontario

Prepared for
Canada Centre for Mineral and Energy Technology
555 Booth Street
Ottawa, Ontario
K1A 0G1

June 1997

AQUATIC EFFECTS TECHNOLOGY EVALUATION PROGRAM

Notice to Readers

Technical Evaluation on Methods for Benthic Invertebrate Data Analysis and Interpretation

The Aquatic Effects Technology Evaluation (AETE) program was established to review appropriate technologies for assessing the impacts of mine effluents on the aquatic environment. AETE is a cooperative program between the Canadian mining industry, several federal government departments and a number of provincial governments; it is coordinated by the Canada Centre for Mineral and Energy Technology (CANMET). The program was designed to be of direct benefit to the industry, and to government. Through technical and field evaluations, it identified cost-effective technologies to meet environmental monitoring requirements. The program included three main areas: acute and sublethal toxicity testing, biological monitoring in receiving waters, and water and sediment monitoring.

The technical evaluations were conducted to document certain tools selected by AETE members, and to provide the rationale for doing a field evaluation of the tools or provide specific guidance on field application of a method. In some cases, the technical evaluations included a go/no go recommendation that AETE takes into consideration before a field evaluation of a given method is conducted.

The technical evaluations were published although they do not necessarily reflect the views of the participants in the AETE Program. The technical evaluations should be considered as working documents rather than comprehensive literature reviews. The purpose of the technical evaluations focused on specific monitoring tools. AETE committee members would like to stress that no one single tool can provide all the information required for a full understanding of environmental effects in the aquatic environment.

The objective of this project was to critically review the recent literature on statistical analysis and interpretation of benthos data with respect to biomonitoring, and to recommend analytical approaches that are robust, objective, effective, and ecologically relevant for monitoring Canadian metal mines. Some participants felt that the scope of this report was limited mostly to methods developed and used in North American freshwater studies. A review of other methods and approaches which have been successfully applied is provided in a discussion paper entitled "*Review of potentially applicable approaches to benthic data analysis and interpretation*" (AETE Report #2.1.3a - Dr. R. Green, March 1999) prepared as a complement of information to this report.

For more information on the monitoring techniques, the results from their field application and the final recommendations from the program, please consult the AETE Synthesis Report to be published in the spring of 1999.

Any comments concerning the content of this report should be directed to:

Geneviève Béchard
Manager, Metals and the Environment Program
Mining and Mineral Sciences Laboratories - CANMET
Room 330, 555 Booth Street, Ottawa, Ontario, K1A 0G1
Tel.: (613) 992-2489 Fax: (613) 992-5172
E-mail: gbechard@nrcan.gc.ca

PROGRAMME D'ÉVALUATION DES TECHNIQUES DE MESURE D'IMPACTS EN MILIEU AQUATIQUE

Avis aux lecteurs

Évaluation technique des méthodes d'analyse et d'interprétation des données sur les invertébrés benthiques

Le Programme d'évaluation des techniques de mesure d'impacts en milieu aquatique (ÉTIMA) visait à évaluer les différentes méthodes de surveillance des effets des effluents miniers sur les écosystèmes aquatiques. Il est le fruit d'une collaboration entre l'industrie minière du Canada, plusieurs ministères fédéraux et un certain nombre de ministères provinciaux. Sa coordination relève du Centre canadien de la technologie des minéraux et de l'énergie (CANMET). Le programme était conçu pour bénéficier directement aux entreprises minières ainsi qu'aux gouvernements. Par des évaluations techniques et des études de terrain, il a permis d'évaluer et de déterminer, dans une perspective coût-efficacité, les techniques qui permettent de respecter les exigences en matière de surveillance de l'environnement. Le programme comportait les trois grands volets suivants : évaluation de la toxicité aiguë et sublétale, surveillance des effets biologiques des effluents miniers en eaux réceptrices, et surveillance de la qualité de l'eau et des sédiments.

Les évaluations techniques ont été menées dans le but de documenter certains outils de surveillance sélectionnés par les membres d'ÉTIMA et de fournir une justification pour l'évaluation sur le terrain de ces outils ou de fournir des lignes directrices quant à leur application sur le terrain. Dans certains cas, les évaluations techniques pourraient inclure des recommandations relatives à la pertinence d'effectuer une évaluation de terrain que les membres d'ÉTIMA prennent en considération.

Les évaluations techniques sont publiées bien qu'elles ne reflètent pas nécessairement toujours l'opinion des membres d'ÉTIMA. Les évaluations techniques devraient être considérées comme des documents de travail plutôt que des revues de littérature complètes.

Les évaluations techniques visent à documenter des outils particuliers de surveillance. Toutefois, les membres d'ÉTIMA tiennent à souligner que tout outil devrait être utilisé conjointement avec d'autres pour permettre d'obtenir l'information requise pour la compréhension intégrale des impacts environnementaux en milieu aquatique.

L'objectif de cette évaluation était d'examiner la documentation récente sur l'analyse statistique et l'interprétation des données sur le benthos aux fins de la surveillance biologique et de recommander des approches analytiques robustes, objectives, efficaces et pertinentes sur le plan écologique, pour la surveillance des mines de métaux canadiennes. Certains des participants étaient d'avis que la portée de ce rapport était limitée en ce qu'il touchait principalement les méthodes développées et utilisées pour des études en eaux douces menées en Amérique du Nord.

Une étude d'autres méthodes et approches qui ont été appliquées avec succès a été publiée dans un document de travail intitulé « *Review of potentially applicable approaches to benthic data analysis and interpretation* » (ETIMA, Rapport n° 2.1.3a - R. Green, mars1999). Cette étude a été préparée pour compléter l'information contenue dans le présent rapport. »

Pour des renseignements sur l'ensemble des outils de surveillance, les résultats de leur application sur le terrain et les recommandations finales du programme, veuillez consulter le Rapport de synthèse ÉTIMA qui sera publié au printemps 1999.

Les personnes intéressées à faire des commentaires concernant le contenu de ce rapport sont invitées à communiquer avec M^{me} Geneviève Béchard à l'adresse suivante :

Geneviève Béchard
Gestionnaire, Programme des métaux et de l'environnement
Laboratoires des mines et des sciences minérales - CANMET
Pièce 330, 555, rue Booth, Ottawa (Ontario), K1A 0G1
Tél.: (613) 992-2489 / Fax : (613) 992-5172
Courriel : gbechard@nrcan.gc.ca

Executive Summary

Scope

The Aquatic Effects Technology Evaluation (AETE) program commissioned a technical evaluation of methods in benthic invertebrate data analysis and interpretation for biological monitoring at mine sites. The objective of the technical evaluation was to review the recent literature and recommend analytical approaches that are valid, objective, effective, and ecologically relevant for monitoring Canadian metals mines. The best analytical methods are those that derive the most useful information and provide the greatest sensitivity in a biomonitoring program at the lowest cost. *Sensitivity*, the ability to detect small or moderate changes in benthic invertebrate community structure against a background of natural spatial and temporal variability, is especially important in a biomonitoring program because sensitive methods can act as early warning systems of impending ecosystem damage, and are more likely to detect subtle effects of chronic, low-level metals loadings.

The technical evaluation covers statistical analysis and ecological interpretation of quantitative data on benthic invertebrate densities derived from more or less simultaneous samples allotted according to a simple spatial design. This study design is based on replicate samples collected at one or more reference sites upstream from, or otherwise outside, the zone of influence of the effluent outfall and at a series of exposed sites downstream. Replication may be either by multiple samples at individual sites or by multiple sites, each sampled once, within larger zones.

Analytical Approach

The favoured statistical approach rests on the premise that a biomonitoring study is essentially a test of a hypothesis, specifically, that a mine is exerting biological effects on a particular water body at a particular time. The investigator begins with a null hypothesis that the mine effluent has no effect, and tests the hypothesis by comparing exposed sites against unaffected reference sites, while attempting to minimize, by careful attention to design and analysis, the possibility of a site difference occurring for reasons unrelated to mining. The conclusion that a mine effect is or is not present is based on *strong inference* because in a spatial design the possibility of another source of downstream effects can never be completely eliminated. Thus, in most routine surveys the pollution source is implicated if the nature and spatial distribution of effects on the benthos are congruent with beforehand expectations based on the nature of the effluent, and there are no other disturbances present to which the effects could reasonably be attributed.

Analysis of Variance (ANOVA) or its derivatives (ANCOVA, MANOVA) is the preferred method of testing for significant differences in species abundances or community metrics among sites in a biomonitoring study. Descriptive multivariate methods such as ordination and clustering may be useful to reduce the complexity of the data set or reveal major patterns, but are not sufficient by themselves to determine an effluent effect. It is only on the basis of statistical tests of hypotheses that a statement can be made, with known probability of error, that the mine is causing deleterious effects on the exposed water body.

Analysis of Covariance (ANCOVA) is a powerful means of reducing variability from habitat variables unrelated to the mine and thereby increasing the sensitivity of the analysis, and careful use of ANCOVA is to be encouraged. Multivariate Analysis of Variance (MANOVA) is preferred over simple ANOVA because it considers several variables at once and thereby reduces the risk of a Type I error (finding a difference where none exists), especially when the variables are correlated. However, the use of MANOVA is restricted in routine monitoring studies by the requirement for large numbers of replicates. Modifications to sampling programs (collection of habitat data at each sampling point, increasing replication and decreasing sample size) that would facilitate the use of these two methods should be promoted.

Simple graphs of species abundances, richness or other variables against sites and distances from point sources are a straightforward and easily comprehended means of presenting benthic invertebrate data. Means and ranges or standard deviations should be included on the graphs along with an indication of statistically significant differences. Large-scale site descriptors such as canopy cover or land use, that cannot be statistically compared can be included on graphs to illustrate broader differences among sites. Graphs from ordinations or clustering dendrograms can also be informative but should not displace simple scatterplots of the original data as the mainstay of data presentation.

The determination of effects of mines is strongest when it is based on the composite results for many taxa and community variables combined in a *weight-of-evidence* argument. The thrust of this approach is to search for trends in taxa densities that are consistent with a hypothesized effect of the effluent or other disturbance. Results for any one taxon alone are not sufficient to reject the null hypothesis, but similar changes in other taxa are taken as confirmation that the observed site difference is real. Hence, this approach uses the weight-of-evidence based on the number and kinds of taxa showing differences between sites, and the strength of the response from each.

Choice of Response Variable

Abundances of common taxa, aggregated into groups of similar organisms if numbers are low, constitute the keystone of the weight-of-evidence analysis for site differences. Individual species or genera are the most varied and sensitive indicators of environmental conditions. The parallel analysis of several taxa provides both an opportunity to confirm the direction of observed trends, and, when combined with knowledge of the biology of the organisms, provide valuable insight into the nature of the stresses affecting the community. Higher taxonomic levels such as insect orders should only be used where lower taxa are too rare or too variable to be useful and the members of the higher taxon are reasonably similar in ecological requirements. To avoid a large number of redundant or unhelpful analyses, it is important to screen the raw data carefully and retain only those variables that are likely to show a statistically significant trend that is consistent with the expected effect of the disturbance. However, all taxa contribute to the weight-of-evidence argument, including those that do not differ among sites.

Selected summary statistics ought to be included in the analysis also, to provide a measure of the severity of effects on the community as a whole. Total abundance of all organisms and total number of taxa per sample are useful and well-established variables but may be unresponsive to slight degradation. *Similarity indices* should be included in site comparisons because: (1) they summarize the overall difference in community structure between reference and exposed sites as a single number; (2) they require no pre-conceived assumptions about the nature of a healthy community; and (3) they can only vary in one direction, avoiding the interpretive problems that arise from stimulation. The most reliable similarity indices appear to be the Bray-Curtis Index and the Per Cent Similarity Index.

Diversity indices, such as the Shannon-Weaver Index, have been popular in pollution assessment, but they tend to be unresponsive to slight or moderate disturbance, especially when it does not involve organic enrichment, and are not recommended for biomonitoring at Canadian mine sites. *Biotic indices* assess water quality based on the presence or absence of indicator species of known tolerance and summarize conditions in a single number. Biotic indices should only be included in biomonitoring at mines when they are applicable to the geographic region and there is reason to expect organic or mixed effluents. Biotic indices must be calculated for each sample and subjected to statistical analysis in the same manner as other variables.

Functional feeding groups are guilds of invertebrate taxa that obtain food in similar ways, regardless of taxonomic affinities. Ecological studies on flowing waters suggest that the proportion of different feeding groups will change in response to disturbances that affect the food base of the system, thereby offering a means of assessing disruption of ecosystem function. The utility of functional feeding groups as variables to estimate impairment of benthic communities at mine sites is uncertain. Research to date has laid too much emphasis on evaluating effects of severe impairment. More research is needed to test the sensitivity and reliability of feeding groups at moderately contaminated sites where the food base has not been directly altered.

Rapid assessment procedures are intended for quick, qualitative assessments of water quality based on preliminary sampling and are not an adequate tool for biomonitoring at mines. Nevertheless, some metrics used in rapid assessment procedures may also be useful in quantitative biomonitoring, and research comparing the sensitivity and accuracy of different metrics should not be disregarded. However, the "multi-metric" approach to biomonitoring, in which a diversity of unrelated metrics is combined into a single number to rank sites, is not sound biologically or statistically. All metrics based on ratios between two variables should also be avoided.

Power

Statistical power is the probability that a test will report a difference between two treatments when they are truly different; it is the statistical analogue of sensitivity. Power is a key element of sound experimental design in biomonitoring that has not been afforded the attention that it deserves. Power analysis should be routinely incorporated into every biomonitoring study. During study design, power should be calculated based on preliminary sampling or data from previous years to ensure that sampling intensity is sufficient to ensure a reasonable probability of detecting site differences of a magnitude deemed to be ecologically significant. Power calculations should also be done on every analysis of variance that fails to detect a significant differences among sites. The power analysis should either demonstrate that the power of the test was reasonable, or determine the magnitude of difference between sites that would be required for a test of reasonable power.

Research Needs

More research on the effects of mine wastes on benthic invertebrates in lakes and rivers, especially their responses to low-level, chronic loading and to mixed metal-organic wastes, would help investigators attempting to formulate hypotheses of expected mine effects. Research to determine the occurrence and significance of stimulation responses at slightly contaminated sites is needed because of the complexity of interpretation introduced by bi-directional responses to disturbance. Experiments to establish toxicity of various metals to a variety of common benthic species in Canadian water bodies would also be useful. All raw data from each biomonitoring study should be archived in a safe, organized, accessible data base for future studies of temporal trends, and possible integration into a network of regional reference sites.

Sommaire

Portée

Le Programme d'évaluation des techniques de mesure d'impacts en milieu aquatique (ETIMA) a commandé une évaluation technique des méthodes d'analyse et d'interprétation des données sur les invertébrés benthiques aux fins de la surveillance des effets biologiques des effluents miniers. Cette évaluation avait pour objectif d'examiner la documentation récente sur le sujet et de recommander des méthodes d'analyse à la fois valides, objectives, efficaces et pertinentes au plan écologique, en vue de la surveillance des effets biologiques des effluents rejetées par les mines de métaux canadiennes. Les meilleures méthodes d'analyse sont celles qui fournissent les informations les plus utiles au moindre coût tout en se révélant les plus sensibles dans un contexte de surveillance biologique. La *sensibilité*, c'est-à-dire la capacité d'une méthode de détecter les changements d'amplitude faible ou modérée dans la structure des communautés d'invertébrés benthiques parmi un ensemble de fluctuations spatio-temporelles naturelles, est une caractéristique particulièrement souhaitable en surveillance biologique, car les méthodes sensibles, en plus de jouer le rôle d'un système d'alerte rapide en signalant la survenue imminente de dommages écosystémiques, sont également les plus susceptibles de détecter les effets subtils des faibles charges chroniques de métaux.

L'évaluation technique portait sur l'analyse statistique et l'interprétation écologique des densités d'invertébrés benthiques estimées à partir d'échantillons prélevés plus ou moins simultanément selon un plan spatial simple. Le plan de l'étude prévoyait la collecte d'échantillons répétés à un ou plusieurs sites de référence répartis en amont ou à l'extérieur de la zone exposée à l'effluent ainsi qu'à une série de sites exposés situés en aval de la zone de rejet. La répétition pouvait être réalisée de deux façons, soit par prélèvement d'échantillons multiples dans des sites individuels, soit par prélèvement d'échantillons uniques dans des sites multiples répartis dans des zones plus vastes.

Approche analytique

L'approche statistique privilégiée repose sur la prémisse selon laquelle une étude de surveillance biologique est essentiellement un test d'hypothèse visant à établir si une exploitation minière a des effets biologiques sur un plan d'eau donné à un moment donné. L'évaluateur aborde le processus analytique en supposant que l'effluent minier n'a aucun effet. C'est l'hypothèse nulle. Il vérifie la validité de cette hypothèse en comparant les sites exposés aux sites de référence non touchés, tout en s'efforçant de réduire le plus possible, en portant une attention particulière au plan de l'étude et à l'analyse, la probabilité que les différences observées à un site donné soient dues à des causes ne présentant aucun lien avec l'activité minière. La part d'inférence dans la conclusion selon laquelle l'activité minière exerce ou non un effet est forte, car dans un plan d'étude spatial, il est impossible d'exclure totalement la possibilité qu'une autre source d'effets en aval soit en cause. C'est pourquoi, dans la majorité des relevés courants, la pollution est mise en cause si la nature et la répartition spatiale des effets sur le benthos sont compatibles avec les attentes initiales formulées en considération de la nature de l'effluent et du fait qu'aucune autre perturbation ne permet d'expliquer les effets observés.

L'analyse de variance ou une méthode dérivée (analyse de covariance ou analyse multivariée de variance) est la méthode recommandée pour vérifier l'hypothèse selon laquelle l'abondance des espèces ou d'autres paramètres des communautés diffèrent de façon significative d'un site à l'autre dans la zone de surveillance biologique. D'autres méthodes d'analyse multivariée descriptive (p. ex. ordination ou groupement) permettent également aider à

simplifier la base de données ou à mettre en évidence certaines tendances, mais elles ne permettent pas à elles seules d'établir de façon définitive si l'effluent minier exerce réellement un effet sur les communautés benthiques. C'est uniquement sur la base des conclusions de tests d'hypothèses statistiques qu'il devient possible d'affirmer, avec un risque d'erreur connu, qu'une exploitation minière a des effets néfastes sur le plan ou le cours d'eau exposé.

L'analyse de covariance contribue à accroître la sensibilité de l'analyse parce qu'elle permet de réduire la variabilité imputable aux variables environnementales ne présentant aucun lien avec la mine. C'est pourquoi son utilisation est fortement recommandée. L'analyse multivariée de variance doit également être préférée à l'analyse de variance conventionnelle, car elle tient compte en même temps de l'effet de plusieurs variables, réduisant ainsi le risque d'erreur de type I (conclure à tort à l'existence d'une différence), en particulier lorsque les variables sont corrélées. Toutefois, le recours à l'analyse multivariée de variance n'est envisageable que dans le cadre d'études de surveillance courante en raison du nombre élevé d'observations répétées requises. La modification des programmes d'échantillonnage (collecte de données sur l'habitat à chaque point d'échantillonnage, augmentation du nombre d'échantillons répétés et réduction de la taille des échantillons) en vue de faciliter l'utilisation de ces deux méthodes devrait également être encouragée.

Les graphiques simples illustrant les variations d'abondance des espèces, de la richesse ou d'autres variables en fonction des sites et de la distance à partir des sources ponctuelles constituent une façon à la fois simple et directe de présenter les données sur les invertébrés benthiques. Les moyennes et les intervalles ou les écarts-types associés à chacune des variables présentées sur ces graphiques devraient également être indiqués, ainsi que des indications concernant le seuil de signification des différences relevées. Bien qu'ils ne se prêtent pas à la comparaison statistique, certains macrodescripteurs des sites comme le couvert végétal ou l'utilisation des terres devraient également figurer sur les graphiques, car ce type d'information permet d'illustrer les différences plus larges entre les sites. Les graphiques et les dendrogrammes obtenus à l'aide d'ordinations et de groupements fournissent également des informations utiles, mais ils ne peuvent remplacer les graphiques simples illustrant la répartition des données dans l'espace.

L'appréciation des effets des exploitations minières est d'autant plus solide qu'elle s'appuie sur l'analyse des résultats composites de nombreuses variables concernant les taxons et les communautés et sur l'examen d'éléments concluants. Cette approche a pour objectif d'examiner les densités de divers taxons en vue de cerner des tendances permettant de confirmer l'existence d'un effet ou d'une autre source de perturbation. Si la présence de changements chez un taxon ne suffit pas à rejeter l'hypothèse nulle, l'observation de changements similaires chez d'autres taxons tend à confirmer l'existence de différences significatives entre les sites. En d'autres mots, cette approche tire parti du poids des preuves fondées sur le nombre et le type de taxons présentant des différences d'un site à l'autre et l'ampleur des réactions dans chaque cas.

Choix des variables indiquant une réponse

Les abondances des taxons communs ou de groupes d'organismes similaires (en cas de faible abondance) constituent les éléments de base de l'analyse par poids des preuves des différences relevées entre les sites. Les espèces et les genres sont les indicateurs les plus variés et les plus sensibles des conditions environnementales. L'analyse en parallèle de plusieurs taxons permet de confirmer le sens des tendances observées et, lorsque la biologie des organismes considérés est prise en compte, fournit souvent des indications fort utiles sur la nature des contraintes qui pèsent sur la communauté. L'utilisation de niveaux taxonomiques plus élevés, tels les ordres d'insectes, ne devrait être envisagée que lorsque l'abondance des taxons inférieurs est trop faible (taxons rares) ou trop variable pour être utile et lorsque les exigences écologiques des taxons supérieurs sont raisonnablement similaires. Afin d'éviter la multiplication d'analyses redondantes ou superflues, il est important de trier les données brutes afin de retenir uniquement les variables les plus susceptibles de mettre en évidence une tendance statistiquement significative compatible avec l'effet présumé de la perturbation. Cependant, tous les taxons sont utiles lorsqu'il s'agit d'examiner le poids de la preuve, y compris ceux dont l'abondance ne varie pas d'un site à l'autre.

Il convient également d'inclure des statistiques sommaires choisies dans l'analyse afin de fournir une mesure de l'ampleur des effets sur la communauté prise dans son ensemble. L'abondance totale de tous les organismes et le nombre total de taxons par échantillon sont des variables à la fois utiles et largement reconnues, mais elles peuvent être insensibles à des dégradations de faible amplitude. L'utilisation *d'indices de similarité* dans la comparaison des sites est également souhaitable, car ces indices : 1) résument en un seul chiffre la différence globale de structure des communautés entre les sites témoins et les sites exposés; 2) n'exigent aucune hypothèse de départ concernant la nature d'une communauté saine; 3) ne varient que dans un sens, permettant d'éviter ainsi les problèmes d'interprétation soulevées par la stimulation.

Les *indices de similarité* les plus fiables semblent être les indices de Bay-Curtis et l'indice de similarité proportionnel. Les indices de diversité, comme l'indice de Shannon-Weaver, ont été largement utilisés dans le cadre d'études de pollution, mais ils présentent le désavantage de ne pas répondre à des perturbations légères à modérées, en particulier en l'absence d'enrichissement organique. C'est pourquoi leur utilisation pour la surveillance des effets biologiques de l'activité des mines canadiennes sur les écosystèmes aquatiques n'est pas recommandée. Les indices biotiques permettent d'évaluer la qualité de l'eau d'après la présence ou l'absence d'espèces indicatrices présentant un seuil de tolérance connue et résument les conditions en un seul nombre. Le recours aux indices biotiques ne devrait être envisagé dans le cadre de programmes de surveillance des effets biologiques de l'activité minière que lorsqu'ils s'appliquent à la région géographique étudiée et qu'il y a de bonnes raisons de soupçonner la présence d'effluents organiques ou mélangés. Ces indices doivent être calculés pour chaque échantillon, et ils doivent être soumis aux analyses statistiques de la même manière que les autres variables.

Les *groupes d'alimentation fonctionnels* sont des guildes formées de taxons d'invertébrés qui obtiennent leur nourriture au moyen de stratégies similaires, sans égard aux affinités taxonomiques. Des études écologiques sur les eaux vives portent à croire que la proportion de groupes d'alimentation différents varie en réponse aux perturbations qui agissent sur la disponibilité des ressources alimentaires à l'intérieur du système. Ces fluctuations permettent donc d'évaluer dans quelle mesure le fonctionnement d'un écosystème est perturbé. L'utilité réelle des groupes d'alimentation fonctionnels comme variables permettant d'estimer le degré de dégradation des communautés benthiques dans les sites miniers demeure à déterminer. Les études menées à ce jour ont accordé trop d'attention à l'évaluation des effets associés à des perturbations importantes. D'autres recherches s'imposent pour évaluer la sensibilité et la fiabilité des groupes d'alimentation dans les sites modérément contaminés où la disponibilité des ressources alimentaires n'est pas trop compromise.

Les *méthodes d'évaluation rapide* permettent d'évaluer rapidement et de façon qualitative la qualité de l'eau d'après les résultats d'un échantillonnage préliminaire et, dès lors, ne sont pas recommandées pour la surveillance biologique des effets de l'activité minière sur les écosystèmes aquatiques. Néanmoins, certains paramètres associés à ces méthodes peuvent également fournir des indications utiles en surveillance biologique quantitative, et les recherches comparant la sensibilité et la précision de divers paramètres ne doivent pas être négligées. Toutefois, l'application de l'approche « multi-métrique » à la surveillance biologique, c'est-à-dire d'une approche qui prévoit le regroupement d'une série de paramètres indépendants en un nombre unique pour évaluer les sites, n'est pas justifiée, ni au plan biologique, ni au plan statistique. L'utilisation de paramètres résultant d'un ratio entre deux variables est également à éviter.

Puissance

La puissance statistique correspond à la probabilité qu'une épreuve révèle une différence entre deux traitements réellement différents. C'est l'équivalent statistique de la sensibilité. Bien qu'elle soit un élément indispensable de tout bon protocole expérimental appliqué à la surveillance biologique, la puissance n'a pas reçu à ce jour toute l'attention qu'elle mérite. L'analyse de puissance devrait être incluse automatiquement dans tout projet de surveillance biologique. Durant la conception de l'étude, la puissance devrait être estimée à partir des résultats d'un échantillonnage préliminaire ou de données recueillies au cours d'années antérieures. Cette façon de faire permettrait de s'assurer que l'échantillonnage possède l'intensité voulue pour faire ressortir des différences écologiquement significatives entre les sites. Des calculs de la puissance devaient également être effectués chaque fois qu'une analyse de variance ne révèle aucune différence significative entre les sites. L'analyse de puissance devrait soit établir que le test est suffisamment puissant, soit déterminer l'ampleur de la différence entre les sites requise pour un test présentant une puissance raisonnable.

Besoins en recherche

Il convient d'entreprendre d'autres recherches sur les effets des effluents miniers sur les invertébrés benthiques des lacs et des rivières, en particulier sur les réponses de ces organismes à de faibles charges chroniques et à des effluents mélangés de nature métallique et organique. Ces travaux aideraient les chercheurs à formuler des hypothèses concernant les effets présumés des mines. Il faut également réaliser des recherches sur l'occurrence et l'importance des réponses aux stimuli dans les sites faiblement contaminés en raison de la complexité de l'interprétation introduite par les réponses bidirectionnelles aux perturbations. Des expériences permettant d'établir la toxicité de divers métaux pour un certain nombre d'espèces d'invertébrés benthiques communes dans les eaux canadiennes pourraient également fournir des données utiles. Toutes les données brutes recueillies dans le cadre de chaque projet de surveillance biologique devraient être versées dans une base de données sure, structurée et accessible en prévision d'études futures sur les tendances temporelles et d'une intégration éventuelle dans un réseau de sites de référence régionaux.

Acknowledgements

This report was written for the Aquatic Effects Technology Evaluation (AETE) program, a co-operative program among the Canadian mining industry and responsible agencies of the federal government and most provincial governments. The project was directed by Lise Trudel of the Canada Centre for Mineral and Energy Technology (CANMET). Thorough and helpful reviews of an earlier draft by Roger H. Green (Dept. of Zoology, University of Western Ontario, London), William Duncan (Cominco Ltd., Trail, B.C.), Glen Watson (Inco Ltd., Copper Cliff, Ont.), and Bill Keller (Ontario Ministry of Northern Development and Mines, Sudbury) considerably improved the final document. The authors are grateful as well to Zsolt Kovats (Golder Associates Ltd., Calgary, Alberta), Dennis Farara (Beak International Inc., Brampton, Ont.) and Richard Pope (Tarandus Associates Ltd., Brampton, Ont.) for providing information and sharing their expertise on the subject of benthic invertebrate analysis. Valuable direction and support were also received from Derek Riehm (Teck Corporation, Vancouver, B.C.), Diane Campbell (CANMET, Ottawa) and the other members of the AETE Technical Committee.

TABLE OF CONTENTS

Page

1. Introduction 1

2. Analytical Framework 4
   2.1 Objectives 4
   2.2 Experimental Design 4
   2.3 Sensitivity 7
   2.4 Power 8
   2.5 Statistical Methods for Hypothesis Testing 9
   2.6 Role of Inference 11
   2.7 The Weight-of-Evidence Approach 12
   2.8 Interpreting Site Differences 14

3. Choice of Response Variables 16
   3.1 Total Density and Species Richness 16
   3.2 Common Taxa 19
   3.3 Diversity Indices 21
   3.4 Biotic Indices 24
   3.5 Rapid Assessment Indices 27
   3.6 Similarity Indices 35
      3.6.1 General 35
      3.6.2 Coefficient of Community Loss 40
      3.6.3 Per Cent Model Affinity 42
   3.7 Functional Feeding Groups 43
      3.7.1 Theory and Practice 43
      3.7.2 Statistical Considerations 48

4. Statistical Methods 52
   4.1 Basic Approach 52
   4.2 Alternatives to the Standard ANOVA 53
   4.3 Improving the Sensitivity of the Basic Approach 55
   4.4 Limitations of the Basic Approach 57
   4.5 Ecological *versus* Statistical Significance 60

5. Interpretation of Statistics 64
   5.1 Inferring Cause and Effect 64
   5.2 Complications 66
   5.3 Effects of Mines 69
   5.4 Benthic Invertebrates and Stream Processes 73

6. Conclusions and Recommendations 77
   6.1 Conclusions 77

6.2 Recommendations 78

7. Literature Cited 81

LIST OF TABLES

Table 1 Typical responses of benthic invertebrate abundance and species richness to various types of stress 17

Table 2 Metrics used or proposed for inclusion in rapid assessment procedures 29

Table 3 Distribution of invertebrate functional feeding groups among four classes of water quality (defined by polar ordination on species distributions) in a Maine river historically polluted with pulp mill effluent 47

Table 4 Change in relative abundances of various taxa of benthic insects in outdoor artificial streams dosed with copper (25 µg/L) or copper (12 µg/L), cadmium (1.1 µg/L) and zinc (110 µg/L) for ten days 72

LIST OF FIGURES

Figure 1 Standard design for environmental assessment of a mine site 5

Figure 2 Example of graphical use of similarity indices to illustrate overall change in benthic community structure in a river receiving wastewater 36

Figure 3 Proportions of functional feeding groups in the benthos at six sites along a Montana river (Clark Fork) contaminated with heavy metals 49

Figure 4 The value of MANOVA 54

Figure 5 Graphical comparison of number of taxa between reference and exposed sites 56

Figure 6 The utility of covariates 58

Figure 7 Example of delayed stimulation from effluents containing both toxins and organic material 67

Figure 8 Survival rates of late-instar nymphs of different species of mayfly exposed to dissolved copper in laboratory streams 70

Figure 9 Effect of insecticide treatment on organic matter processing in a small Appalachian stream 75

1. Introduction

The Aquatic Effects Technology Evaluation (AETE) program was established to review appropriate technologies for assessing the effects of mine effluents on aquatic ecosystems. AETE is a co-operative program among the Canadian mining industry, several federal government departments and eight provincial governments; it is co-ordinated by the Canada Centre for Mineral and Energy Technology (CANMET). The program has two stated objectives: to help the Canadian mining industry meet its obligations for environmental effects monitoring in the most cost-efficient manner; and to evaluate new and established monitoring technologies that could be used for assessment of environmental effects of mining. The program is designed to be of direct benefit both to the industry and to government by evaluating and identifying cost-effective technologies to meet environmental monitoring requirements. The program includes three main areas: acute and sublethal toxicity testing, biological monitoring in receiving waters, and water and sediment monitoring.

The AETE program includes field evaluations of biological monitoring technologies to be used by the mining industry and regulatory agencies to assess the effects of mine effluents on aquatic ecosystems. The goal of the program is to recommend specific methods, or groups of methods, that will permit accurate characterization of environmental effects in the receiving waters in as cost-effective a manner as possible. A pilot field test was conducted in 1995 to fine-tune the study approach. In 1996, preliminary surveys were carried out at several mine sites across Canada. The field evaluation of selected monitoring methods will then take place at five of these mine sites in 1997.

Community structure of benthic macro-invertebrates, the insects, worms, molluscs and other organisms living on the bottoms of rivers and lakes, is included in the field study as an indicator of environmental quality and mine effluent effects. The success of a biomonitoring program depends not only on the methods used to collect, sort and identify organisms in benthic samples, but also on the statistical methods applied to the resulting data and the ecological interpretations given those data. Realizing the importance of data analysis in benthic invertebrate biomonitoring, and the wide assortment of methods available, the Technical Committee decided that a technical evaluation of methods in benthic invertebrate data analysis and interpretation should be carried out as part of the AETE program. The objective of the technical evaluation is to review the recent literature on statistical analysis and interpretation of benthos data with respect to biomonitoring and to recommend analytical approaches that are valid, objective, effective, and ecologically relevant for monitoring Canadian metals mines. The best methods are those that derive the most useful information and provide the greatest sensitivity in a biomonitoring program at the lowest cost.

The earliest use of benthic invertebrates in biomonitoring was for assessments of gross organic pollution from sewage outfalls, which produces large and predictable effects on downstream fauna. Since then, methods for benthic invertebrate monitoring have been extended to cover most kinds of disturbance, and ecologists and statisticians have devoted considerable attention to extracting more and better information from benthos data. The result has been a diverse, but too often confusing, panoply of new methods from which the investigator must hope to choose the best. Different methods have advocates and different approaches have adherents, and the eagerness of each point of view to convince the others engenders lively debate. Many advances in data analysis are best applied to specific circumstances or objectives; there is no one best analytical approach for all purposes. In particular, techniques designed for studies of aquatic biology or for regional surveys of water quality, where trends and patterns are being explored, are not necessarily the best for local biomonitoring, where a specific hypothesis of directed environmental change is being tested (Green and Montagna 1996). Fortunately, there is also now a sizable literature of tests and comparisons among different analytical methods, from which the critical reader may derive some guidance on how best to proceed.

This report discusses the statistical analysis and biological interpretation of quantitative data from a benthic invertebrate biomonitoring program. It assumes invertebrate densities have been determined in replicate, quantitative samples from a sequence of sites along a river or a more complex equivalent in a lake, presumably following a discharge plume. Study design and field methods such as choice of sampling sites and sampling gear, and laboratory procedures such as how invertebrates are sorted and counted are discussed in a companion report (Taylor 1997) and for the most part are not considered here. However, most statistical methods depend upon a certain sampling regime or assume samples have been collected in a particular way, so some discussion of the sampling implications of various analytical choices is unavoidable.

The objectives of this report are very specific. It is not intended as a how-to manual for benthic invertebrate studies, but rather as a critical review of some options for improving the cost-efficiency and effectiveness of biomonitoring for the mining industry. The emphasis is on increasing the sensitivity and reliability of analysis without sacrificing statistical rigour. The primary goal is to point out alternatives to conventional approaches, and to search out methods that are robust, objective, effective and ecologically relevant for monitoring Canadian metals mines. The report provides a short list of recommended methods along with a scientific rationale for each and an evaluation of its strengths and limitations. Data gaps where more research is needed and methods that have promise but require more testing before they can be confidently used are also identified.

The discussion is directed toward the kind and magnitude of environmental effects to be expected from wastewater discharges from metals mines in Canada, and therefore assumes the potential effects would arise mostly from heavy metals and inorganic sediments contained in wastewaters, and to a lesser extent organic pollution or nutrient enrichment. The text is unavoidably weighted toward monitoring of flowing waters, in particular small to medium-sized streams and rivers, because that is where the majority of waste effluents have traditionally been discharged and hence where the majority of research has been done. However, analysis of data from more complex, two-dimensional sampling arrays such as arise from sampling of large rivers, lakes and the oceans is also included.

The literature review was limited to works published since 1980. Particular attention has been devoted to studies from the last ten years, because many new ideas have been proposed in that period that have potential to substantially improve data analysis and interpretation. Biological monitoring in general has been reviewed by contributors in Loeb and Spacie (1994) and a comprehensive review of biomonitoring with benthic invertebrates, covering many of the topics in this report, was published by Rosenberg and Resh (1993). Their work provides an expert synthesis of the literature up to about 1990. Relying on the work of Rosenberg and Resh to cover the older literature, the effort in this work was concentrated on (a) updating the literature review to include accounts from 1990 through 1996, and (b) re-examining the very general evaluation of Rosenberg and Resh (1993) in the narrower context of methods for the mining industry.

2. Analytical Framework

2.1 Objectives

Before any detailed discussion of analytical methods can begin, it is necessary to look at the specific goals of biomonitoring at mine sites, and see how those goals affect the approach taken to analysis and interpretation. Biomonitoring practised for regulatory or environmental protection purposes at mines is intended only to assess the effects of effluents or run-off on the local aquatic environment. It is not intended as a comprehensive survey of the aquatic community nor as a broader survey of water quality in the region. This kind of monitoring is usually undertaken annually or biennially along a water body potentially disturbed by wastewaters from a mine site either as process effluents from the tailings pond or as diffuse runoff from disturbed land or tailings. These routine assessments have three primary objectives:

(1) to determine whether the mine is having a detrimental effect on the benthic invertebrate community within the water body;

(2) to measure the nature and severity of any detrimental effects; and

(3) to determine how far away from the mine the effects extend.

2.2 Experimental Design

This report is concerned with analysis of data sets based on one-time sampling without consideration of temporal trends (although analysis of changes in environmental quality through time may be a long-term goal). Hence, it assumes more or less simultaneous sampling at all sites, and that sampling effort was allotted according to a spatial design, with one or more reference (or control) sites upstream of the effluent outfall or otherwise outside its zone of influence, and a series of potentially affected sites downstream from the effluent outfall or within its zone of influence (Anderson 1990, Klemm *et al.* 1990). It further assumes the collection of multiple replicates at each site (Figure 1). For convenience, the term "upstream" is used in this report to include all reference or control sites unaffected by the mine effluent, and "downstream" refers to exposed sites potentially influenced by the mine, whether in standing or flowing waters

5

A more recent variant of the simple spatial design is to define several sites within larger zones (as shown in Figure 1 as the Reference, High Exposure and Low Exposure Zones) and collect a single sample at each site. Sites then become the replicates for statistical comparisons of zones (Cuff and Coleman 1979). This approach is used in the Environmental Effects Monitoring program for Canadian pulp and paper mills and is discussed in more detail in Taylor (1997). As long as a single sample is used to define each site, statistical analysis of this variant is identical to the classic design; in essence the site replicates from the classic design have merely been spread over a larger area. Any reference to replicates in the present text should be construed to mean sites within zones if this variation of the design is being considered.

The simple spatial design, also known as the Control-Impact design, is perhaps the weakest from a purely statistical perspective because there is built-in confounding between the effect of the effluent and ordinary variation between points in a lake or river (Hurlbert 1984). Careful site selection and simultaneous measurements of important habitat variables are essential to the success of the design (Camacho and Vascotto 1991). Designs incorporating repeated sampling before and after a disturbance begins are superior to the simple spatial design (Green 1979, 1989) but cannot be applied to routine monitoring at mines already in operation. The consensus among practitioners of biomonitoring in Canada is that the spatial design is satisfactory for most benthic invertebrate studies examining effects of point sources if appropriate modifications are made regarding the number and location of sampling sites (Environment Canada 1993).

In contrast to the simple Control-Impact design for biomonitoring studies, the *reference condition* (or reference areas) approach entails comparing benthic invertebrate communities at potentially affected sites against a variety of reference sites in the same physiographic region. The benthic invertebrate communities at the reference sites are taken to represent the normal condition, unaffected by human influence, and the nature and degree of impairment at affected sites is determined by how far their benthic communities depart from those at reference sites.

The reference condition approach merges with the classic sampling design as the number of control sites increases; hence the two approaches represent opposite ends of a continuum. Recommendations on sampling design are beyond the scope of this report; a previous review (Taylor

1997) concluded that it was advantageous to have more than one control site, but that the reference condition approach was not presently workable for biomonitoring at mines. The review comprising this report assumes a classic sampling design; ideas derived from the reference condition approach are included if they are broadly applicable.

Routine biomonitoring using a spatial design is intended to assess environmental conditions at the moment the samples were taken. Analysis of trends in water quality through time are not the main objective of these programs. Nevertheless, the industry or regulators may want to know if there has been a deterioration of conditions over time, or whether changes in wastewater treatment or plant operations have improved effluent quality. Thus, it is helpful if the biomonitoring program uses consistent methods of sample collection and data analysis from one year to the next, so that re-analysis of the accumulated data for temporal trends is possible.

Cumulative data analysis requires the return of high quality, accessible, organized data from each study. Where monitoring studies are undertaken by different investigators in different years, data incompatibility and loss are major hindrances to long-term studies. An organized effort is called for to ensure that the raw data from each study, after appropriate quality checks, are included in a central data base for each mine or region, and that those data are maintained and made available to other researchers. The data submitted should include counts of organisms (on a square metre basis) from all samples, not just site means and variances, along with background information on the exact location of each site and the sampling methods used. As recommended in Taylor (1997), a reference collection of benthic invertebrates should also be maintained for every mine site and be made available to consultants or researchers when each biomonitoring study is undertaken.

2.3 Sensitivity

Analytical methods for data from biomonitoring studies should be statistically rigorous and biologically meaningful. Within those bounds, the key attribute of the best methods is *sensitivity*, the ability to detect relatively small changes in the benthic community. Sensitivity implies distinguishing real site differences caused by subtle disturbances from random differences caused by natural variation. A sensitive method has a high "signal to noise ratio", where spatial patterns in the benthos caused by mining constitute the signal, and natural place-to-place variation is the noise.

Sensitivity is crucial to effective biomonitoring at mines because in the absence of spills or accidents, effects of mines on downstream water bodies will often be quite small. Detecting chronic, low-dose exposures to contaminants is a more important issue for biomonitoring at these sites than detecting acute, high-dose exposures. This is not to say that severe impairment does not occur; on the contrary, detection of spills or other major upsets is a valid and common use of benthic invertebrate monitoring. But biological responses to strong disturbances are relatively easy to detect. It hardly matters whether sampling and analytical methods are sensitive when the impairment of the benthic community is large and conspicuous.

However, there is only limited value in biomonitoring to document the effects of severe disturbance, which is obvious in any case. It is far more useful to use biomonitoring as an early warning system, to signal ominous changes in environmental quality at their inception, so that corrective action may be taken before major environmental damage occurs (Bunn 1995, Humphrey *et al.* 1995). This latter use of biomonitoring is the intent of routine surveys at mines and industrial sites where the quality of aquatic ecosystems is monitored as a normal part of environmental vigilance. Statistical techniques that work well for strong pollution may be less suitable to monitoring the smaller and more subtle effects to be expected from low-level metals contamination or other disruptions associated with mining. Therefore, the selection of optimal analytical methods must include sensitivity as an important criterion.

The sensitivity of many new or modified methods suggested in the literature is often hard to judge because field tests of such methods tend to use large geographic scales, often entire river systems (e.g., Faith 1990, Barton and Metcalfe-Smith 1992, Palmer *et al.* 1996), or choose test sites that are severely compromised by strong pollution or some other major disturbance (e.g., Winner *et al.* 1980, Rabeni *et al.* 1985, Barbour *et al.* 1992, Novak and Bode 1992, Wallace *et al.* 1996). The latter is particularly true of rapid assessment methods (Section 3.5) and biotic indices (Section 3.4), which are intended to gauge biological effects over the full range of pollution gradients. A comparison of an index between pristine sites and severely disturbed sites is perhaps seen as necessary to establish its spectrum of response, but it does nothing to confirm whether it works well on less drastically altered communities. Consequently, many of the methods received with enthusiasm for biomonitoring in other regions do not end the quest for sensitive methods for monitoring at Canadian mines.

2.4 Power

The statistical analogue of sensitivity is *power*, the probability that a test will report a difference between two treatments when they are truly different. Power is discussed in Section 4, and cogently reviewed by Fairweather (1991). Traditionally, researchers developing new methods in biomonitoring have paid scant attention to statistical power, concentrating instead on ensuring that tests were accurate and unbiased. It is equally important to find methods that are powerful, because failing to detect a real impairment can have serious consequences for the environment (Faith *et al.* 1991, Peterson 1993).

The power of a test depends on, among other things, the magnitude of the difference between means that is to be detected, and the variance in the data. Ideally, the design of the study would incorporate information from previous work or preliminary sampling, and sampling intensity would be calibrated to detect differences of known magnitude in selected variables (Bernstein and Zalinski 1983, Maher and Norris 1990, Camacho and Vascotto 1991). In practice, logistical and budget restraints often compromise the sensitivity of biomonitoring programs. Nevertheless, the investigator should at least be aware of the statistical power of the data, and hence what magnitude of differences among sites can be detected when they are indeed present.

Power analysis is seldom explicitly done in biomonitoring studies, but knowing the power of statistical tests is crucial to the credibility of the study. Failing to reject the null hypothesis of no impairment in the exposed zone is not equivalent to demonstrating that the hypothesis is true if the power of the test is low (Peterman 1990a). In that circumstance, it is possible that an effect did exist but could not be detected by the statistical test used, given the sampling intensity and population variance. Power calculations can be used to differentiate between null hypotheses that probably are true and those that require more thorough sampling for a convincing test (Peterman 1990b). Hyland *et al.* (1994) provide an excellent example of power analysis in the context of a marine environmental study.

2.5 Statistical Methods for Hypothesis Testing

A biomonitoring study is essentially a test of a hypothesis, namely, that there is not an effect of a mine on a particular water body at a particular time. This point is critical to the whole approach to data analysis. The investigator begins with a null hypothesis that the mine effluent has no effect, and tests the hypothesis by comparing control or reference sites (upstream or outside of the influence of the mine) against exposed sites (downstream or within the influence of the mine), while attempting to minimize, by careful attention to design and analysis, the possibility of a site difference occurring for reasons unrelated to mining. It is only on the basis of statistical tests of hypotheses that a statement can be made, with known probability of error, that the mine is causing deleterious effects on the adjacent water body.

It follows that statistical methods, conventional or novel, that test for differences between reference and exposed populations should be the cornerstone of data analysis in biomonitoring surveys. The many powerful techniques for detecting pattern in animal assemblages (cluster analysis, ordinations and their ilk) have their place, as exploratory tools to examine faunistic similarities among sites and to uncover spatial structure within the data set. Ordinations and similar multivariate methods are popular in aquatic ecology, and their value as tools for pollution assessments has been frequently advocated (Marchant *et al.* 1984, Warwick and Clarke 1991, 1993, Agard *et al.* 1993, Warwick 1993, Norris 1995). But as these methods can only indirectly detect significant differences among sites, they are inadequate to support a decisive analysis of biomonitoring data unless assisted by more direct hypothesis-testing methods.

Descriptive multivariate methods are most appropriate for preliminary analysis when the investigator has limited knowledge of the system and wants to generate hypotheses for subsequent testing (Fore *et al.* 1996). In this application, an ordination or similar technique would be used to reduce the complexity of the data set by replacing a large number of variables (invertebrate taxa), many of them potentially intercorrelated, with a manageable number of uncorrelated variables, such as site scores from the ordination or dominant groups from a cluster analysis. The reduced variable set would then be used in a multivariate or univariate analysis of variance to test for differences among the sites.

While the popularity of this analytical approach cannot be denied, its appropriateness for biomonitoring is open to question. Fundamentally, in a biomonitoring study there is no need to search for structure in the data because the data are deliberately structured to begin with, by the placement of sampling sites relative to point sources. Ordinations and similar multivariate methods describe the total data structure among all the individual observations (samples). There is thus no guarantee that the methods will detect differences between upstream and downstream benthic communities even when such differences are profound. Moreover, many attempted ordinations yield no useful reduction in dimensionality or produce uninterpretable axes.

Some investigators prefer to use the two-stage approach of data reduction by a multivariate method followed by hypothesis tests on the reduced variable set. Our contention remains that this approach is not the most effective for routine biomonitoring studies. It is sufficiently challenging to assess the ecological significance of changes in simple variables, such as a 50% reduction in the density of a genus of mayfly; how is one to decide important effect sizes in a derived variable like the second axis of a correspondence analysis? Whether or not this step is included, however, the hypothesis-testing step is critical; otherwise the basic question of whether the mine is or is not affecting benthic communities downstream is not objectively tested. Therefore, ordinations or similar multivariate techniques should not be presented by themselves without support from hypothesis-testing statistics.

The evaluation of methods in the present work is based on the hypothesis-testing approach to biomonitoring. Within that framework the review examines the choice of variables with which essential attributes of the benthic community may be quantified, and the statistical methods available to compare them among sites. Simple graphical methods are advocated for data presentation, and inferential statistical methods rather than descriptive multivariate methods are used to relate mining to biological effects.

2.6 Role of Inference

Interpretation of benthic invertebrate community structure at exposed sites in a biomonitoring program is based on the principle of strong inference, in which detrimental effects are defined operationally as any significant variance from the community structure at a comparable control site, usually upstream (Underwood 1991, DFO & Environment Canada 1995), although environmental quality may sometimes be judged against expected community composition for the region. In a simple spatial design, however, a finding of a significant difference between sites above and below a point source cannot be construed as definitive proof that the effluent or other disturbance is responsible for the change without further supporting evidence. Thus, in most routine surveys the pollution source is implicated as the cause of observed impairment if: (1) the zone of impairment begins below the suspected source with a consistent pattern of recovery farther away and there are no other sources nearby to which the impairment could reasonably be attributed; and (2) the effects on the benthic community are congruent with expectations based on the nature of the effluent or disturbance.

In the simplest case, the nature of influences from a mine may be known or can be predicted based on literature, background site information and effluent chemistry. In more comprehensive studies, toxicity tests, plume delineations and other ancillary information may be available to help confirm effluent effects (Van Hassel *et al.* 1988). Concordance between expected and observed effects bolsters the inference that the mine is responsible for observed impairment of the benthic community. The key point is that the inference must be supported by a reasonable, beforehand expectation based on independent evidence that the point source would cause the effect observed. Section 5 discusses this issue in more detail.

2.7 The Weight-of-Evidence Approach

A biomonitoring field survey typically produces a wealth of raw data, with numerous species at each site and a very wide range of abundances among sites and taxa. Investigators are faced with the dilemma of selecting or deriving from the raw data a workable number of variables for comparison among sites, a step often referred to as *data reduction*. In practice there is a broad continuum in the extent to which the raw data are compressed or filtered for statistical treatment. At one end of the spectrum is what might be termed the *summary statistic* approach, and at the other, the *weight-of-evidence* approach.

The summary statistic approach selects a few key variables to characterize a community or defines a new variable that expresses overall community structure. Species richness and total density of all organisms are popular choices, as are numbers of species or individuals in certain sensitive groups, such as the Ephemeroptera and Plecoptera. Biotic indices, discussed in Section 3.4, and similarity indices (Section 3.6) would also be included in this class.

The argument for summary statistics is that, if they are well chosen or derived, then one or a few figures can effectively capture the behaviour of the entire community and allow site differences to be easily resolved. Summary statistics promote a clear, straightforward analysis and are easy to present to non-specialists. Statistics such as similarity indices include information from every species in the assemblage, and can be weighted to emphasize abundances of rare or common species. Biotic indices also include information from all taxa and consider the relative sensitivity of each to determine a single-number ranking for the site. The observation that disturbed communities tend to lose species and gain or lose individuals is the rationale behind simple variables like total density or species richness.

Two compelling arguments can be raised against the summary statistic approach: first that it is insensitive; and second that it is insufficient. Summary statistics are undeniably useful, and they will demonstrate effects of pollution or other disturbances strong enough to markedly alter community structure. However, coarse measurements such as total density, or number of species (or even species richness within one group, such as the popular Ephemeroptera-Plecoptera-Trichoptera (EPT) index) cannot be expected to signal slight or subtle changes in community structure (see Section 3.5).

Moreover, these coarse variables are insufficient by themselves because they say nothing of the nature of the community change. If total density declines by 50%, did all species decline or just some? Which groups lost most species, and what did those species have in common? Were they all sensitive to metals, or to sediments? What changes, exactly, resulted in a decline in community similarity with respect to upstream sites, or a decrease in a biotic index? It is necessary to establish which taxa or groups were responsible for observed site differences, to confirm and understand the stress on the system. Summary statistics are discussed in Section 3, and some of them are very powerful. But their limitations must be realized.

Many of these objections to summary statistics are answered by multivariate methods. The various methods for ordination separate sites along gradients defined by the benthos itself, so they presumably reflect the dominant environmental features that produced the distribution of organisms. Many of these methods also indicate which taxa were most influential in determining the axes. However, as discussed earlier (Section 2.5), multivariate methods for pattern analysis are not the optimum choice for use in biomonitoring unless they facilitate tests of hypotheses. Multivariate ANOVA, on the other hand, can be used to test for site differences using several variables simultaneously, without the loss of information inherent in a summary statistic. Unfortunately, use of MANOVA in biomonitoring studies is severely restricted by sample size requirements (see Section 4). If field methods switched to higher numbers of smaller samples, as recommended previously (Taylor 1997), this restriction would be considerably relieved.

In contrast to the summary statistic approach, the weight-of-evidence approach is based on parallel analysis of several to many variables, especially densities of individual taxa. The potential number of variables is limited only by the number of taxa collected; summary statistics of various kinds can also be included, but are never used alone. The thrust of this approach is to search for trends in taxa densities that are consistent with a hypothesized effect of the effluent or other disturbance. Given random chance and the considerable number of statistical tests, results for any one taxon might be misleading. Additional taxa, indicating a parallel change, are taken as confirmation that the observed site difference is real. Hence, this approach uses the weight-of-evidence based on the number and kinds of taxa showing differences between sites, and the strength of the response from each.

The weight-of-evidence approach can be considered an extension of the summary statistic approach, and at least partly overcomes the limitations of sensitivity and sufficiency described earlier. This approach is more sensitive because it can detect changes in individual taxa, whether or not there is a significant change in the structure of the whole community. It can also reveal the nature of the disturbance through the degree to which different taxa are affected. The progress of recovery away from the point source may also be seen if some species recover sooner (i.e., nearer the source of disturbance) than others.

The effectiveness of the weight-of-evidence approach still depends upon a thoughtful choice of variables for comparison, but if the data are examined thoroughly, subtle changes like a species replacement are unlikely to go unnoticed. The weight-of-evidence approach requires a detailed, painstaking look at the whole community and cannot be done by rote. There is no substitute for understanding the ecology of the system.

At the extreme, and if applied without judgement, this approach can be laborious, redundant, inefficient and confusing. With too many variables there is a risk of becoming overwhelmed by the analysis and losing the important trends in a cloud of statistics. High variability in individual taxa densities can make differences among sites difficult to detect. It is therefore important to screen the raw data carefully and retain only those variables that are likely to show a statistically significant trend that is consistent with the expected effect of the disturbance.

While it is clear that the injudicious use of either the summary statistic approach or the weight-of-evidence approach should be avoided, in routine biomonitoring studies there is a strong tendency for data to be analyzed too incompletely (not considering enough variables) or too coarsely (not using sensitive approaches) to detect subtle changes in the ecosystem under study. An astonishing number of industry studies rely overwhelmingly on just three variables: total density, species richness and some measure of overall community structure. This approach is both ineffective and inefficient: the former because real effects may go undetected or incompletely understood; the latter because a great deal of effort has been expended collecting, identifying and enumerating the organisms in the samples. Having made that investment, it is not cost-effective to simplify the analysis by comparing just a few coarse variables. The conclusion therefore, is that a reasonable weight-of-evidence approach will be most effective for biomonitoring at mine sites, especially where slight to moderate disturbance is anticipated.

2.8 Interpreting Site Differences

Biologists realize that statistics are merely tools to help identify patterns and differences in the benthic fauna of their study sites. They cannot by themselves determine the ecological or cultural significance of those differences. Statistics should be used to make an objective decision about the presence of a change downstream, but other means must be used to determine whether the change is significant ecologically and whether it is consistent with the expected effects of the point source or disturbance. The ecological significance of changes can only be determined by examining the nature and the degree of change to impaired communities in the context of natural variation and disturbances from other sources, both natural and anthropogenic.

The biological interpretation of changes in community structure is the most important and difficult step in the analysis, and the one for which it is most difficult to provide firm guidance. It is important to realize, however, that while statistical analysis is important, solid and sensitive results can often be obtained with well-known methods without recourse to novelties. Stewart-Oaten *et al.* (1992) and Fore *et al.* (1996) point out that, while decisions concerning site differences must have a statistical basis, it is far more important to interpret the magnitude and ecological significance of population changes than to fuss over exact estimates of statistical significance. Nevertheless, it is important to know: (1) which variables showed significant differences among sites; (2) which variables were not significantly different in tests of reasonably high power; and (3) which variables could not be expected to show changes of a magnitude perceived to be important because the power of the test was low. Only with all this information can a comprehensive, weight-of-evidence evaluation of the effects of the mine on the receiving ecosystem be made.

3. Choice of Response Variables

3.1 Total Density and Species Richness

The total abundance of all species and the total number of taxa in the sample are the simplest and easiest variables to obtain from a set of benthic samples, and are frequently the first variables (sometimes the only variables) compared among sites. Methods guides almost universally recommend using total abundance and species richness (properly taxon richness if all specimens are not identified to species) as key indicators of environmental stress (Klemm et al. 1990, Anderson 1990, Beak 1990). These two variables are fundamental attributes of community structure, and respond in broadly predictable ways to most types of environmental stress (Table 1). Changes in either variable have direct biological implications and are easily compared among sites. Therefore it makes sense to include abundances of taxa and individuals in the analysis.

The problem apparent in the literature is not with inappropriate use of total abundance variables, but in over-reliance on these two simple sums. While analysis of total abundance and taxon richness may prove useful, and should be retained, they are usually insufficient of themselves to form the basis for a sensitive biomonitoring analysis of biomonitoring data. Yet in both the scientific literature and industry biomonitoring, analysis too often extends no further than these two variables, sometimes coupled with a simple diversity or biotic index (e.g., Beak 1990, Crunkilton and Duchrow 1991, Battegazzore *et al.* 1992).

The total density of benthic invertebrates will respond in a predictable way to gross perturbations, but it is too coarse a measure to detect subtle trends. A profound re-arrangement of species composition at a site could occur without any marked change in total abundance because increases in some species will be masked by compensatory decreases in others (Norris and Georges 1993). It is only widespread decreases (as in response to toxicity) or disproportionate increases (as by tolerant species in response to enrichment) that will be reflected in total abundance. In addition, total density tends to be very sensitive to habitat characteristics and therefore shows wide natural variation (Klemm *et al.* 1990), although variability can be reduced by careful site selection. Consequently, only large changes in abundance are usually detectable with the level of sampling effort ordinarily employed in biomonitoring studies.

Table 1. Typical responses of benthic invertebrate abundance and species richness to various types of stress. (Source: Klemm *et al.* 1990)

Stress	Effect on Abundance	Effect on Species Richness
Toxic substance	Reduces	Reduces
Severe temperature change	Variable	Reduces
Silt	Reduces	Reduces
Low pH	Reduces	Reduces
Inorganic nutrients	Increases	Variable
Organic enrichment (low dissolved oxygen)	Increases	Reduces
Sludge deposits (non-toxic)	Increases	Reduces

Taxon richness is a more responsive variable than total abundance, but is also most effective for demonstrating strong pollution effects (e.g., Barton and Metcalfe-Smith 1992). Slight or moderate impairment is sometimes expressed by replacements of one species by another and changes in the relative proportions of species rather than by net loss of taxa. Unfortunately, the response of taxon richness to disturbance may not always be monotonic: while gross organic pollution causes a characteristic loss of intolerant species (and great increases in densities of tolerant species), species richness at moderately enriched sites often increases relative to unproductive controls, because new species colonize to take advantage of the abundant food supply (Cook 1976). Species richness declines again when enrichment becomes more severe.

Consistent trends of declining species richness, both within specific insect orders, and in the entire community, have sometimes been observed in response to smooth gradients of toxic metals. Rasmussen and Lindegaard (1988) provide a clear-cut example from a Danish river with a longitudinal gradient of dissolved ferrous iron. At the sites farthest from the source, where iron concentrations were <0.2 mg/L, the river supported 67 species of benthic invertebrates. Species richness decreased closer to the iron source along a tightly linear trend to a low of 10 species at the nearest site (10 mg Fe/L). The same pattern of gradual loss of species could be observed in miniature within higher taxa such as insect orders, or even among species within the mayfly genus *Baetis*. Hence, in this study species richness was a dependable indicator of toxic effects over a gradient from slight to severe impairment. Total abundance, by contrast, was only affected at the highest iron concentrations, as was observed elsewhere in response to copper (Leland *et al.* 1989).

If taxon richness is to be used as a measure of benthos community impairment, the question arises of whether rare species should be included. Elsewhere it has been argued that statistically rare species, those that are represented by only a few individuals at each site, should simply be deleted from the species list before analysis begins (Taylor 1997). The information provided by species at the edge of detection is too variable to aid in distinguishing clean and impaired sites, and removing them would lead to lower richness counts but considerably reduced variances within sites. Many workers hold the opposite view, however (Environment Canada 1993). Fore *et al.* (1996) argue that rare species provide critical clues to biological condition because these species will be eliminated first when a site is perturbed. Perhaps so, but for most statistically rare species, counts per sample are so low at reference sites that it is impossible to distinguish, with a reasonable sampling effort, whether the absence of those species at an exposed site is an effect of the effluent or merely a sampling artifact. Clearly, this question can only be resolved by examining field data.

Even if total abundance and taxon richness indicate statistically significant differences among sites in a study, a meaningful weight-of-evidence analysis should go further toward understanding the response of the benthic invertebrate community. For example, if abundance is declining, are all species declining or just a few common species? If taxon richness is declining, which species disappear first? Do they share a common taxonomic group, habitat, feeding mode or sensitivity to some particular stress (e.g., sediments)? Answers to these questions can reveal much about the nature and severity of the stress acting on the community, and allow more informed judgements about the seriousness of the impairment and the need for remedial action.

3.2 Common Taxa

Much insight into the nature of responses to pollution or disturbance can be gained by comparing abundances of individual taxa among sites. This step is integral to the basic idea of biomonitoring, namely that the distribution and abundance of benthic organisms is a reflection of the environmental influences visited upon them (Johnson *et al.* 1993). Comparisons among sites for individual taxa are also at the heart of a weight-of-evidence approach to analysis. The utility of simple comparisons of individual taxa should not be underestimated; Hellawell (1977) found that comparison of taxon abundances among sites was more useful than total abundance, species richness, or a wide variety of biotic indices, diversity indices, and similarity indices for measuring pollution effects in two English rivers.

Either univariate (ANOVA, etc.) or multivariate (MANOVA) methods can be used for comparisons (Section 4), and most practising biologists are well aware of the exigencies of each approach (Environment Canada 1993). The key question here is how to choose taxa for detailed comparisons from the long list provided by most sampling programs. Species that are rare at all sites may be set aside because no statistically meaningful differences will be detected with those species. Among the remainder, the number of tests can be reduced if the data are examined graphically, and only those taxa showing an apparent difference in mean density between sites, and that consistent with a hypothesized effluent effect, are subjected to statistical analysis.

Univariate comparisons of each common taxon may not be the most statistically efficient approach. The natural variability in abundance of individual species can make it difficult to quantify changes consistently and precisely (Norris and Georges 1993). If site differences are small or absolute numbers of organisms in a particular taxon are low, sensitivity can be improved by combining closely related taxa for analysis. For example, numbers in several species of *Baetis* might be combined and analysis done on the entire genus. This method sacrifices information for each individual species to achieve a more powerful test for the higher taxon. It should only be done for phylogenetically and ecologically similar species, and will only be successful if all combined species show approximately the same response to the disturbance.

A more sophisticated solution is to use multivariate analysis of variance on several taxa at one time. MANOVA has the advantage of being a more powerful test; conversely it is much more complex and demanding to use than ANOVA, and has more restrictions as well. In particular, the total number of observations in a MANOVA must be greater than the number of variables (taxa) plus the number of cells (sites, times, or site-time combinations) or there will be no error degrees of freedom for conducting tests (Smith *et al.* 1990). Numbers of observations in excess of five times the number of variables are frequently recommended for reliable hypothesis testing with MANOVA (Environment Canada 1993). These restrictions provide a powerful impetus for keeping the number of variables small.

Combining taxa for analysis, whatever the method used, should be done circumspectly and attempt to parcel together taxa that are ecologically similar or which display a parallel distribution pattern among sites. Collapsing species into genera will reduce sensitivity but often reveals the same main trends (Hellawell 1977). Comparisons of whole insect orders (Plecoptera, Ephemeroptera, Coleoptera etc.) among sites are frequently uninformative and suffer from a weak ecological basis. Members of a common order, especially in such large groups as Trichoptera and Diptera, may be very different with respect to habits, habitat, and sensitivity; therefore, a substantial shift in community structure can take place without seriously affecting the proportions of insect orders, especially at sites that are already dominated by one group (Hellawell 1977). A recent workshop of consultants and researchers in biomonitoring (Environment Canada 1993) suggested the following guidelines for when to use higher taxa:

(1) no information on tolerance or sensitivity is available for lower taxonomic levels, or

(2) all or most taxa at lower levels are similar ecologically relative to the differences among higher taxa, or

(3) abundances of all taxa within a higher taxon are positively correlated.

3.3 Diversity Indices

It has long been known that in benthic invertebrate communities, as in most other animal assemblages, there are typically a few very abundant species, a number of less abundant but common species, and a large number of species represented by only a very few individuals. The use of indices of community diversity in water quality monitoring is based on the concept that the structure of benthic communities (that is, the relative numbers of abundant, common and rare species) may be changed by perturbations of the environment because some species will be suppressed more than others by the perturbation. It follows that the degree of change in community structure will reflect the intensity of the environmental stress (Hellawell 1977).

Diversity as originally defined was seen as the mathematical analogue of variety, and therefore had two components: *species richness*, the number of species in the community, and *evenness*, the number of individuals in each. Theoretical maximum evenness is reached when every species in the community is represented by the same number of individuals. Some measurements also incorporate abundance (Metcalfe 1989). This definition of diversity contrasts with the modern term *biodiversity*, which is usually taken as equivalent to species richness. Based on the perception that diversity was an important issue in community dynamics, ecologists devoted considerable attention to its measurement and there are now many formulas from which to choose (see Washington 1984 for a thorough review). The most popular among these are Simpson's diversity index (Simpson 1949) (also known as Simpson's Dominance Index) and the Shannon-Weaver Diversity Index (Shannon and Weaver 1949), often represented by the symbol H'.

The popularity of diversity indices in water quality assessment arose from two sources. The first was the hypothesis, promulgated in the works of MacArthur (1955), Margalef (1968) and Odum (1969) that there was a relationship between species diversity and ecosystem stability, defined as the capacity of an ecosystem to resist perturbation. To calculate diversity of a sample of benthic invertebrates was therefore to measure a fundamental attribute of ecosystem structure and to gain direct insight into the system's biological integrity (Washington 1984). Because gross organic pollution from sewage discharges characteristically caused a loss of species intolerant of low oxygen tensions and a proliferation of tolerant species, diversity indices offered an attractive and simple way to assess community changes that appeared to have a valid scientific basis. Wilhm and Dorris (1968) formally proposed using the Shannon-Weaver Index to assess effects of organic pollution on aquatic communities, and even went so far as to offer a quantitative scale, with pollution being indicated by values of H' less than 3.

The attractions of simplicity of calculation (the species do not need to be identified) and straightforward interpretation have made the Shannon-Weaver index consistently popular since its introduction, and it soon migrated into widespread and routine use for pollution assessments. Advances in theoretical and field ecology, however, have since shown that the relationship between diversity and stability is not so simple, and the theoretical basis of the Shannon-Weaver Index is at best questionable (Hurlbert 1971, Goodman 1975, Washington 1984).

More to the point, a long line of empirical tests have repeatedly demonstrated that the Shannon-Weaver Index is woefully insensitive to many community responses to stress (Cook 1976, Perkins 1983, Chadwick and Canton 1984, Taylor and Roff 1986, Shaeffer and Perry 1986, Pontasch and Brusven 1988, Ferraro *et al.* 1989, Pontasch *et al.* 1989, Boyle *et al.* 1990, Battegazzore *et al.* 1992, etc.) Diversity is particularly insensitive to effects of toxins, including heavy metals, because these often reduce densities of all species more or less equally, so that diversity is unaffected. Chadwick and Canton (1984) tested the Shannon-Weaver Index and three other diversity indices in a Colorado river receiving zinc, iron and cadmium loading from abandoned mine works. Although there was a significant and obvious decline in abundance of all species below the mines, none of the diversity indices responded to the change because species richness was not sharply reduced. Similarly, Peckarsky and Cook (1981) measured effects of metal-contaminated acid mine drainage by examining colonization of artificial substrates above and below an active metal mine. Toxicity of the drainage was conspicuous both in the sharply lower densities of common species in downstream cages (by a factor of four) and in the number of animals found dead. Yet the Shannon-Weaver Index actually increased downstream. The lower total density and reductions in common species increased the evenness of the community, leading to a higher H'.

In addition to reports of poor performance of the Shannon-Weaver Index at mine sites, the most directly relevant for the Canadian mining industry, this and other diversity indices have been found inadequate or insensitive to effects of copper (Perkins 1983), fuel spills (Pontasch and Brusven 1988), agriculture (Barton and Metcalfe-Smith 1992) and complex industrial/municipal effluents (Pontasch *et al.* 1989, Battegazzore *et al.* 1992). Even in the domain for which it was originally proposed, organic enrichment, the performance of the Shannon-Weaver Index is disappointing. Cook (1976) examined benthic biota along 16 km of a slow-flowing, soft-bottomed stream in New York receiving mild enrichment from residences and farmland. She found no correlation between the Shannon-Weaver Index and biochemical oxygen demand (BOD), dissolved oxygen or counts of coliform bacteria -- all the classic indicators of organic pollution -- although a simple biotic index (Chandler's score) showed a close correlation with all three. Cook labelled H' "a very imprecise, if not dubious, pollution index."

Finally, even when the diversity index does respond to community changes, it is far too insensitive to monitor slight to moderate pollution. For example, Jones *et al.* (1981) showed that clean to slightly enriched streams in Missouri could be ranked according to their pollution status using a biotic index, but by the Shannon Weaver Index all streams were designated unstressed. Species diversity is not even a strict function of impairment; small nutrient additions or enhancements of the food base may cause an increase in abundance without excluding species, with the result that the diversity index goes up (Cook 1976). Even slight metal toxicity may increase diversity. Additions of copper to artificial streams naturally colonized by benthic invertebrates depressed H' except at the lowest concentration, where it increased slightly. Low-level copper toxicity selectively removed rare and very abundant species first, leading to an increase in evenness (Perkins 1983). While in this example the Shannon-Weaver Index was accurately reflecting changes in the invertebrate community, it would still be inadequate for biomonitoring because either an increase or a decline in the index could indicate copper toxicity.

In addition to the telling evidence of insensitivity, diversity indices have been heavily criticized on pragmatic and statistical grounds (see Metcalfe (1989) for a complete list). For examples, the sampling distribution of H' is poorly known, which precludes comparisons based on parametric statistics; indices are sensitive to sample size and taxonomic resolution; and many aquatic ecosystems have naturally low diversity in the absence of disturbance. These shortcomings are serious of themselves, but are less compelling than field evidence of the inadequacy of these measures.

Diversity indices have been widely used and extensively tested. Many studies have now demonstrated that diversity indices in general, and the Shannon-Weaver Index in particular, are faulty in theory and unreliable in practice. Even under the best circumstances the Shannon-Weaver Index is demonstrably far too insensitive to be useful for modern biomonitoring. The emerging consensus is that the diversity index will only reliably detect changes that are so conspicuous that the index is unnecessary anyway.

Although they have not seen the extensive use accorded the Shannon-Weaver Index, other indices of diversity such as Simpson's Index offer little improvement. Washington (1984) concluded from his thoroughgoing review that none of the extant indices were satisfactory and their ecological foundation, uncertain to begin with, had collapsed. New indices continue to be proposed (Osborne *et al.* 1980, Camargo 1992a, Smith and Wilson 1996) and tested (Beisel *et al.* 1996) and the concept of diversity may still be useful in ecological studies, but there is only limited value in this parameter for biomonitoring. The conclusion here is that diversity indices are not good candidates for examining small or moderate effects of disturbance from Canadian metals mines.

3.4 Biotic Indices

A biotic index is a way to express the biological condition of a water body, based on its benthic invertebrate community, as a single number. Biotic indices are based on the concept of the *indicator species*: that certain species are fastidious with respect to environmental conditions they will tolerate, and therefore the absence of those species means that those conditions have not been obtained. The degree of pollution of an aquatic system can be classified by quantifying the variety and abundances of species present with different tolerances or sensitivities to a particular pollutant. Most biotic indices were originally designed for organic pollution (sewage) and most are applicable only in running waters, although lake indices are also available (Johnson *et al.* 1993). Washington (1984) provides a detailed review of many biotic indices and Metcalfe (1989) reviews their history and use in Europe.

Biotic indices have always been more popular in Europe and the United Kingdom than in North America. The most recent versions across the Atlantic are the Belgian Biotic Index in Belgium (De Pauw and Vanhooren 1983) the Indice Biologique Global in France (AFNOR 1985) and the Modified Biological Monitoring Working Party (BMWP) Score in U.K. (Armitage *et al.* 1983). Historically, all these indices arose from the Trent Biotic Index, developed in 1960 for the Trent River Authority (Woodiwiss 1964). A later version of that index (Chandler's Score) was adapted for South Africa by Chutter (1972). Chutter's Index in turn formed the basis for Hilsenhoff's biotic index applicable to central North America (Hilsenhoff 1977, 1987). A more recent index by Lenat (1993) for the southeastern United States has the same structure as Hilsenhoff's index.

Although they vary widely in details and complexity, all modern biotic indices have essentially the same structure. Quantitative or qualitative benthos samples are collected from the site of interest and the index is calculated in four steps:

(1) A *tolerance value*, usually ranging from 1-10, for each species or higher taxon is read from a list of values for biota of the region. Some species may not have tolerance values and are not included in the index.

(2) The tolerance value is multiplied by the abundance of the taxon in the sample, either directly or according to an abundance classification.

(3) The products from step 2 are summed across all taxa.

(4) The total from step 3 is divided by the number of taxa in the sample to derive the biotic index.

The final index value provides a summary of the biological condition in the water body, which can be judged against a scale of quality created by the designers of the index.

Biotic indices have been extensively tested and their strengths and shortcomings are quite well known. The modern versions have undergone repeated revision and appear to be effective, if used sensibly, for their original purpose of assessing organic pollution against an absolute scale of severity. Their utility as tools for biomonitoring at mine sites, however, is at best very limited.

First, the applicability of biotic indices designed for organic pollution to inorganic contaminants such as nutrients, metals and suspended sediments or to physical disruption from siltation or dewatering is dubious. There may be broad correlations in sensitivity at least among the major insect orders (e.g., mayflies appear to be one of the most sensitive groups to many stresses) but the correspondence for lower taxa is weak (Norris and Georges 1993). Hence, conventional indices would be inaccurate if applied without modification.

Development of new indices specifically for mine wastes would be a significant undertaking. Clements *et al.* (1992) have made the first attempt to define a biotic index specifically for heavy metals, the Index of Community Sensitivity. The most promising attribute of this index is that tolerances are defined objectively, according to survival in artificial streams dosed with heavy metals at representative field concentrations. The test animals are drawn from artificial substrates and all the colonizing species are tested at once. In field tests the Index of Community Sensitivity worked well at sites near where the invertebrates were tested, but was less reliable when applied further away.

This idea is promising, but serious limitations to its general use remain. The tolerance values are based on one acute exposure to a single metal (copper), and responses to chronic loadings or other metals may be different. The tolerance values would need to be re-defined for different regions, and possibly for different metals, although there is some evidence that sensitivities to various heavy metals are similar among benthic insects (Clements *et al.* 1992). Finally, the possibility of multiple stresses from different kinds of contaminants (metals, sediments, treated sewage) cannot be excluded, and a successful index for mixed effluents cannot be guaranteed.

Even if biotic indices were shown to be applicable at some mine sites, they are not designed to be sensitive. Most biotic indices define a range of values from pristine to severely degraded. The index is geared to respond over this entire range and consequently may not respond well to slight or moderate impairment (Frutiger 1985). Most tests of biotic indices in the literature deliberately use steep pollution gradients (Rabeni *et al.* 1985, Barton and Metcalfe-Smith 1992, Wallace *et al.* 1996) where the success of the index is judged by its accuracy at differentiating more or less pristine sites from sites that have been gravely perturbed. This is significant because the response of many indices to disturbance is not linear: a relatively small change to a simple, impoverished community in very polluted water signals a large increase in the index, whereas a change of the same magnitude at a clean site has much less effect, given the large number of other species already present (Rabeni *et al.* 1985).

Biotic indices are more useful for categorizing sites within a region than for warning of small changes in environmental quality from one site to another. These indices are designed to produce an absolute ranking of water quality on a universal scale that is intended to represent the ideal community in an unstressed ecosystem. This universal scale is not set up for comparisons among sites, the intent of a biomonitoring program. The significance of given change in the index (say, from 9.1 to 8.6) from one site to another is difficult to state.

All biotic indices are inherently regional, limited by the geographic ranges of the taxa that define them (Frutiger 1985). No single index can be expected to apply across Canada, given the vast land area and the variety of biophysical regions that it encompasses. Some of the more general indices, in particular Hilsenhoff's Index, have been shown to apply over a reasonably wide area, and even European indices may sometimes be applied with modifications for local fauna. Barton and Metcalfe-Smith (1992) applied the Belgian Biotic Index and Hilsenhoff's Index to a badly polluted river system in Quebec. They expanded Hilsenhoff's Index to include non-arthropods (oligochaetes and molluscs) and modified the Belgian Biotic Index by substituting indigenous species whose tolerances could be estimated.

However, biotic indices would have to be developed at least for each of the major regions if they were to be used country-wide in Canada. The tolerance values used for biotic indices are largely subjective, so the accuracy and effectiveness of regional indices is likely to vary, especially if they are defined on an *ad hoc* basis, as in the example above. Maher and Norris (1990) question whether biotic indices would be worthwhile in Australia on the ground that the work needed to define them would be prohibitive.

Finally, biotic indices should be used as a component of a statistically based analysis of biomonitoring data, not a substitute for it. Norris and Georges (1993) point out that biotic indices are frequently reported without any statistical analysis at all, as if the index, once calculated, is an absolute number without error. Naturally these indices are as prone to variation as any other variable, and sensible analysis must compare differences among sites with variation among samples. This entails calculating the index separately for each replicate sample, rather than from the pooled sample or a qualitative survey, and calculating a mean and standard deviation in the usual manner. The sampling distributions of biotic indices are poorly known (Norris and Georges 1993); Narf *et al.* (1984) and Stark (1993) have proposed simple methods for statistical comparisons, but they have not yet seen wide use. Randomization tests, described in Section 4, could also be used to compare biotic indices. The uncritical interpretation of trends in biotic indices without support from statistical computations is one of the major shortcomings of their present use (Norris and Georges 1993).

If biotic indices are included in a biomonitoring program they must be seen as one indicator of biological conditions, to be interpreted in concert with other evidence. Indices should be seen as a way of condensing data for clarity or for presentation and not as a way of avoiding a rigorous and complete analysis of the data (Brinkhurst 1993, Suter 1993). Biotic indices, like any single-number summary variable, do not provide insights into the workings of the community and the nature of differences among sites, and therefore have few benefits to offer over conventional analysis. Yet the data needed to calculate biotic indices are the same as would be required for more traditional statistics (Norris and Georges 1993) with the addition of interpretative data on pollution tolerances. Hence, considering all the limitations outlined above, there would appear to be no compelling reason to include biotic indices among the tools for analysis of biomonitoring data from mines.

3.5 Rapid Assessment Indices

Rapid assessment approaches are designed to identify water quality problems associated with point-source and nonpoint-source pollution or other anthropogenic perturbations and to document long-term changes in water quality within a region. Rapid assessment procedures sharply reduce the cost associated with a biomonitoring program by using a number of time-saving measures, including qualitative sampling, generic or family-level taxonomy, and standard, simple measures of community composition, termed *metrics* (Resh and Jackson 1993). The results of all the metrics are often combined into a single index that expresses the overall condition of a site (Barbour *et al.* 1996).

Rapid assessment methods have been assessed in a companion report on field methods (Taylor 1997), which concluded that these quick-evaluation methods were neither sensitive enough nor robust enough to be useful for biomonitoring at Canadian mine sites. However, some of the many metrics that have been proposed for inclusion in rapid assessment protocols could still be used for statistically based site comparisons, and research on metrics for rapid assessment approaches can reveal which metrics are sensitive and robust and which are too noisy or redundant (Barbour *et al.* 1992, 1996).

Metrics used in rapid assessment can be arranged in five classes (Table 2): richness measurements (numbers of taxa), enumerations (abundance or proportions of different groups), diversity and similarity indices, biotic indices, and functional feeding groups. The utility of each of these classes is discussed elsewhere in this section. Metrics concerning abundances and proportions of various taxa are likely to be the most useful for detection of minor differences between sites. Species richness within various large groups should also be useful, but may be unresponsive to slight or moderate disturbances that change abundances without eliminating species.

Field tests of rapid assessment approaches have generally compared regional reference (clean) sites against substantially impaired sites, because it is this kind of stress that the indices are designed to monitor (Kerans and Karr 1994, Wallace *et al.* 1996). Others have not included impaired sites at all (Barbour *et al.* 1992). Nevertheless some insights may still be gained from these studies. Resh and Jackson (1993) tested a subset of the metrics in Table 2 on California streams subjected to chronic (thermal pollution) or acute (acid spill) disturbance, or no disturbance. While both false positives and false negatives were common for all metrics, some were more reliable: all richness measures, Margalef's Index, a family-level biotic index and the proportion of the scraper functional group. Three other diversity indices, all methods classed as enumerations (see Table 2) and other functional group metrics were inaccurate when tested on either disturbed or undisturbed sites.

Table 2. Metrics used or proposed for inclusion in rapid assessment procedures. (Sources: Resh and Jackson 1993, Kerans and Karr 1994, Resh *et al.* 1995, Barbour *et al.* 1996, Fore *et al.* 1996).

Category	Metric	Expected Response to Disturbance
Richness Measures	Number of taxa	Decline
	Number of Ephemeroptera, Plecoptera and Trichoptera (EPT) taxa	Decline
	Number of Coleoptera taxa	Decline
	Number of Chironomidae taxa	Decline
	Number of Orthocladiinae taxa	Decline
	Number of Tanytarsini taxa	Decline
	Number of families	Decline
	Number of intolerant snail and mussel taxa	Decline
	Number of Crustacea and Mollusca taxa	Variable
	Number of species in selected genera (*Pteronarcys, Baetis, Ephemerella* etc.)	Decline
Enumerations	Number of individuals (or biomass)	Variable
	Number of Chironomidae individuals	Increase
	% EPT individuals	Decline
	% Chironomidae individuals	Increase
	% Tribe Tanytarsini individuals	Decrease
	% Dominant taxon	Increase
	Relative abundance of different groups (Insect orders, Gastropoda, Isopoda, Oligochaeta etc.)	Variable
	Ratio of Tanytarsini/Chironomidae	Decline
	Ratio of Orthocladiinae/Chironomidae	Increase
	Ratio of EPT/Chironomidae individuals	Decline
	Ratio of Hydropsyche/Trichoptera	Increase
	% Individuals in numerically dominant taxa	Increase

	% Non-Dipterans	Decline

Table 2. Metrics used or proposed for inclusion in rapid assessment procedures (continued)

Category	Metric	Expected Response to Disturbance
Enumerations (Continued)	% of Insects that are not Chironomidae	Decline
	Five dominant taxa in common between two sites	Decline
	Number of intolerant taxa	Decline
	% Tolerant groups	Increase
Community Diversity and Similarity	Shannon-Weaver Diversity Index	Decline
	Margalef's Diversity Index	Decline
	Menhinick's Index	Decline
	Simpson's Dominance Index	Decline
	Coefficient of Community Loss	Increase
	Equitability	Decline
	Jaccard Coefficient	Decline
	Pinkham-Pearson Community Similarity Index	Decline
	Number of dominant taxa in common	Decline
	Number of all taxa in common	Decline
	Quantitative Similarity Index	Decline
	% change in taxa richness between two sites	Increase
	Number of unique species per site	Increase
	Number of missing EPT taxa compared with reference site	Increase
Biotic Indices	Index of Community Integrity	Decline
	Belgian Biotic Index	Decline

	Biotic Condition Index	Decline
	Chutter's Biotic Index	Decline
	Hilsenhoff's Biotic Index	Decline
	BMWP Score	Decline
	Chandler Biotic Score	Decline

Table 2. Metrics used or proposed for inclusion in rapid assessment procedures (continued)

Category	Metric	Expected Response to Disturbance
Biotic Indices (Continued)	Indicator-organism presence	Variable
	ISO Score	Decline
	Community Tolerance Quotient	Increase
	Saprobic Index	Increase
	Dominance of tolerant groups	Increase
	Indicator Assemblage Index	Decline
Functional Measures	% Shredders	Decline
	% Scrapers	Decrease
	% Filterers	Decline
	% Gatherers	Variable
	% Predators	Variable
	% Predators (except flatworms, Chironomidae)	Decline
	% Omnivores and scavengers	Increase
	Number of scraper taxa	Variable
	Number of shredder taxa	Variable
	Number of scraper and piercer taxa	Decline
	Ratio of scrapers/filterers	Decline
	Ratio of Trophic Specialists/Generalists	Decline
	Functional group similarity with reference site	Decline

Barbour *et al.* (1992) compared 17 metrics from the United States Environmental Protection Agency (U.S. EPA) Rapid Bioassessment Protocols (Plafkin *et al.* 1989) using a data base of 110 unperturbed stream sites in Oregon, Colorado and Kentucky. They tested the metrics for variability, consistency within each of eight ecoregions, and ability to distinguish mountain sites from plains sites. Their main conclusions were these:

(1) Taxon richness and the EPT index were both successful but highly correlated, as would be expected.

(2) Ratio indices were all highly variable and therefore had very little power to distinguish among sites. Variability in the functional group ratio scrapers/filterers was reduced by redefining the metric as scrapers/(scrapers+filterers) thereby converting it to a percentage.

(3) The ratio of (*Cricotopus*+*Chironomus*)/Chironomidae was also highly variable across regions, but might still be useful in local studies to detect organic or metals loadings.

(4) Eleven metrics proved robust and reliable: taxon richness, EPT index, Pinkham-Pearson index, quantitative similarity index, Hilsenhoff biotic index, % dominant taxon, five dominants in common, Hydropsychidae/Trichoptera, scrapers/(scrapers+filterers), abundance of shredders and a functional group similarity index.

This study provides a valuable warning about the weakness of ratios. However, the validity of the comparisons has been questioned because no impaired sites were included (Brussock 1993, Resh *et al.* 1995). A similar study in Florida that did include polluted sites in the assessment (Barbour *et al.* 1996), found many of the same metrics were most dependable, as judged by the criteria of variance, sensitivity and redundancy. Besides total taxa and EPT taxa, the contributions of Chironomidae (taxa), Crustacea plus Mollusca (taxa or individuals) and Diptera (individuals) were considered good metrics. The percentage of the dominant taxon, two indices (Shannon-Weaver Index and the Florida Index) and three functional groups (gatherers, filterers and shredders) also made the list. The sites defined as "impaired" for comparison in this study suffered serious degradation, such as dissolved oxygen <2 mg/L or toxic effluent composing >25% of low flow.

The study by Kerans and Karr (1994) in the Tennessee Valley was similar to that of Barbour *et al.* (1992), except that streams were all within one river system and sites affected by a variety of disturbances (forest clearing, agriculture, industrial waste discharges) were included. Fourteen metrics were valuable in discriminating sites, exhibited concordance with other measures of site quality and were relatively uncorrelated among themselves: numbers of taxa (of mayflies, stoneflies, caddisflies, intolerant snails and mussels, and total); relative abundances of oligochaetes, omnivores, filterers, grazers, predators and the clam *Corbicula*; dominance by two most common species; and total density of all invertebrates. Rejected metrics included abundance of Chironomidae and three functional feeding groups that were either insensitive or correlated with better metrics.

The work of Fore *et al.* (1996) differs from the previous studies in that it was aimed at finding metrics to classify streams in Oregon which were mostly disturbed by logging and associated road-building. Once again total numbers, dominance by common species, total taxon richness, and numbers of taxa of Ephemeroptera, Plecoptera and Trichoptera were strong candidates for site discriminations. The other successful metrics were associated with effects of sedimentation, expressed as numbers or taxa of sediment-tolerant organisms. Perhaps surprisingly, none of the nine metrics for functional feeding groups were useful in this study.

It bears repeating that the applicability of all these regionally based studies to biomonitoring at mine sites is limited both by their large geographic scales and the generally strong pollution gradients employed. Within that scope, however, several metrics consistently reappear as reliable indicators of disturbance: total taxon richness, total abundance, dominance, and abundance or taxon richness of mayflies, stoneflies and caddisflies (separately or as an EPT index). Dominance has been expressed variously as the percentage of the most abundant 1-5 species, although Fore *et al.* (1996) claim it is most effective if more than just the single numerically dominant taxon is included. Functional feeding groups are less universally successful, but the grazer group appeared to be the most reliable. Many of the other successful metrics are specific to a particular kind of stress. Finally, ratio metrics should be avoided because the higher variance severely reduces their capacity to distinguish impaired and reference sites.

The work of Poulton *et al.* (1995) on a metal-contaminated river in Montana is one of the few studies to test rapid assessment metrics against effects specific to metals mines. They tested eight metrics for correlation with metals concentrations in samples of benthic invertebrates from this heavily contaminated river. Their results mirror those found in larger studies: taxon richness, EPT richness, taxon richness among Chironomidae and percentage of the dominant taxon were all reliable; ratio metrics (EPT/Chironomidae, scrapers/filterers) were poor and highly variable; and the Hilsenhoff Biotic Index was ill-suited to metals pollution.

The success of Chironomid species richness in this study reflects a number of earlier studies that have found chironomids useful indicators of metal pollution because of their numerical abundance and wide range of sensitivities to metals (Waterhouse and Farrell 1985, Armitage and Blackburn 1985). Winner *et al.* (1980) found such a strong correlation between chironomid individuals (positive) or species richness (negative) and metals concentrations in two rivers in Ohio that they proposed % Chironomidae as a quick index of metal contamination. Tests of the index in other, less extremely contaminated systems have found a great deal of variability in the index, especially if organic enrichment is also present. Barton and Metcalfe-Smith (1992) found the index failed to signal metals in the Yamaska River, Quebec, but midges were already very common in that river because of gross organic pollution. Whether the index would be sensitive enough for slight to moderate pollution is questionable, but it merits a closer look.

A common extension of the multimetric approach to rapid assessment is to combine the various metrics into a single summary index, such as the Benthic Index of Biological Integrity (Kerans and Karr 1994, Barbour *et al.* 1996, Fore *et al.* 1996). Proponents argue that the index summarizes "ecosystem integrity" by combining measures of population structure (taxon richness and abundances) with measures of ecological processes (functional feeding groups). However, as Suter (1993) has pointed out, these summary indexes are merely arbitrary, ambiguous combinations of disparate measurements without any rationale for how or why they are combined and no clear idea of what the resulting index means. Such indices have no scientific justification beyond convenience, no objective units or scale, and in fact disguise or obscure the real changes occurring in the affected community. The properties (metrics) composing the index contain the useful information about real properties of benthic communities, and lumping them together into a single number is neither necessary nor scientifically justified.

The rapid assessment approach was designed for quick screening of water bodies and many of the simple, count-based metrics will not be sensitive enough to detect slight or moderate impairment, especially that which does not eliminate species. Still, the more sensitive metrics should be considered during analysis of biomonitoring data at mines. It is important to avoid the temptation to use metrics as an easy way to avoid statistical rigour, by substituting simple sums and visual comparisons of sites. To be useful for biomonitoring at mines, these measures must be included in a solid statistical analysis and metrics that do not lend themselves to statistics should be discarded.

3.6 Similarity Indices

3.6.1 General

A *similarity index* is any one of a variety of simple mathematical functions, usually ranging from 0-1, designed to summarize the concordance in species composition between two species lists, based on numbers or relative abundances of shared species. Hence, a similarity index is basically a measure of the similarity of the structure of two communities (Washington 1984). There are very many similarity indices available, and in the literature the whole group or subset may be known as similarity indices, dissimilarity indices, community comparison indices, correlation coefficients or distance indices. Similarity measures were first developed by plant ecologists to compare vegetation patterns among different locations (Hruby 1987), but their use has since spread to many kinds of community studies, including disturbance effects on freshwater invertebrates (Washington 1984). A change in community composition below an effluent outfall results in a lowering of the similarity between reference and exposed sites.

Some similarity indices, such as Jaccard's coefficient of similarity or Kendall's rank correlation coefficient, use only the presence or absence of taxa in each species list. Quantitative indices consider both presence-absence data and the relative or absolute abundance of each species at each site, and can therefore detect differences between communities that share all the same species. Examples of quantitative similarity indices include the Bray-Curtis index, Morisita's similarity index, squared Euclidean distance and the Pinkham-Pearson index (Pontasch and Brusven 1988).

Similarity indices are used in biomonitoring studies in two ways. A similarity index can be calculated between an upstream control or a local reference site and one or more potentially impaired sites below an effluent discharge or disturbed area. The more common practice is to calculate all possible similarity indices between all pairs of sites to define a similarity (or dissimilarity) matrix, which is then used to order or group the sites using cluster analysis or ordination. While the capacity of these multivariate techniques to sort out complex spatial patterns is indisputable, the utility of a simpler, direct graphical approach supported by inferential statistics should not be disregarded. As illustrated in Figure 2, a graph of similarity indexes against location relative to a point source can illustrate the response and recovery of the benthic communities exposed to an effluent quickly and clearly (Pontasch and Brusven 1988). Faith *et al.* (1995) used the Bray-Curtis index to summarize differences in benthic invertebrate communities above and below a source of uranium mine effluent; they used parametric statistics to establish significant differences, and then ordinated data from different sites and sampling times to look for changes in the disturbance response in different seasons.

Community comparisons using similarity indices are strengthened if there are two or more control or reference sites. The similarity between the reference sites can be taken as an indicator of the degree of natural variation between sites, and thereby define a lower bound on the similarity expected below the point source in the absence of disturbance. For example, if the average index value among reference sites was 0.80, a value of 0.70 between reference and exposed sites could be taken as indicative of unusual stress, and act as a flag prompting either further investigation or remedial action (Figure 2). Of course any quantitative variable could be used in a similar manner (using the variance among reference sites to define the degree of natural variation) and indeed should be in a careful monitoring study. Similarity indices appear to be useful to include in these comparisons because (1) they summarize the response of the entire community, (2) they require no pre-conceived assumptions about the nature of a healthy community, and (3) they can only vary in one direction (see later).

Statistical comparisons of similarity indices among sites are necessary if these are to be included in a rigorous biomonitoring program. ANOVA or *t*-tests are sometimes used for this purpose (e.g., Faith *et al.* 1995), but some argue that ordinary parametric statistics are not really appropriate here because the index values, being each based on data from two sites, are not independent. Also, data from each sample contribute more than once to replicate values of the index, when they are used to calculate an index value with each of the replicates from the other site. Pontasch *et al.* (1989) and Smith *et al.* (1990) suggest a simple procedure based on permutations (equivalent to Mantel's Test) that allows rigorous analysis of similarity index data (see Section 4). The method is straightforward to execute with a computer and its utility deserves to be more widely tested.

Similarity indices digest the information contained in two species lists into a single number. Consequently, the reservations expressed about diversity indices and biotic indices, i.e., that they say nothing of the nature or direction of the change between sites, apply with equal strength to similarity indices. Relying on similarity indices alone as the bulwark of data analysis would result in a superficial and uninsightful study. As with other summary statistics, similarity indices are most effective as one component of a combined analysis relying on a suite of effective variables, including abundances of individual taxa.

Similarity indices differ from other indices, however, in that they define effects of disturbance strictly in terms of the deviation of the downstream communities from those at control sites, without regard to the nature of the change. Hence, the fundamental assumptions underpinning similarity indices are different from those involved in diversity indices or biotic indices, which judge sites according to an absolute scale based on community structure (diversity indices) or presence of intolerant species (biotic indices). Similarity indices implicitly define the control site community as ideal, and record any change from it as a decline in similarity.

This approach has a number of implications. First, similarity indices can only produce a unidirectional response to disturbance, regardless of how downstream communities are affected. The question of interpreting indices that sometimes increase in response to disturbance does not arise, because all such changes cause a decline in similarity. For example, Perkins (1983) found that while copper toxicity tended to depress diversity (H') of artificial stream communities, very low concentrations actually increased it because of selective mortality of dominant species. Similarity indices, on the other hand, all recorded a uniform decline in similarity with respect to control at any copper concentration. A replacement of one species by another in the same tolerance category would not affect a biotic index, but would always change a similarity index.

Similarity indices are therefore free of any assumptions about what constitutes a healthy community, and instead make the weaker assumption that the natural condition, as defined by controls, is the standard, and any deviation from it is significant. On the one hand, this assumption is powerful for monitoring at mine sites because impairment of the benthos from disturbances upstream is already accounted for (in the structure of the reference site communities), and the index measures only the effect of any influences between the reference and exposed site. On the other hand, because any change is seen as significant, these indices do not distinguish between reduced richness or abundance from toxicity and increases from enrichment. Even recovery downstream could be marked by a decline in similarity if the reference sites were already suffering impairment.

A great number of similarity indices have been proposed for various purposes, and each has its particular strength and weaknesses. Different indices do not necessarily agree, and the choice of index can radically alter results of multivariate techniques like cluster analysis (Bloom 1981, Hruby 1987). Washington (1984) deplored the confusing profusion of similarity measures and asked for research to sort out the best. A number of researchers have responded to his call, and it is now possible to make some statement about relative merits of different indices.

Among quantitative indices, many studies agree that the most reliable index, especially where changes in both absolute numbers and species proportions are seen as important, appears to be the Bray-Curtis Index (Bray and Curtis 1957), also known as the Czekanowski Quantitative Index (Hellawell 1977, Perkins 1983, Hruby 1987, Pontasch *et al.* 1989, Pontasch and Brusven 1988, Faith *et al.* 1991).

The Bray-Curtis index provided "the most meaningful condensation of data" out of five indices compared for their capacity to summarize multispecies toxicity tests (Pontasch *et al.* 1989), and it also proved superior for clustering sites along an Australian river (Marchant *et al.* 1984). The Bray-Curtis Index proved superior to seven other indices for monitoring effects of uranium and gold mines, providing consistently high statistical power because of its low temporal variability at control sites and sensitive response to disturbance (Faith *et al.* 1991). Bloom (1981) compared four similarity indices using generated community composition data in which the degree of overlap, from 100% to 10%, was precisely known. The Bray-Curtis index was the only index that accurately tracked the degree of community overlap. Moreover, the Bray-Curtis index was the only index tested that was unaffected by the nature of the communities being compared.

Where only relative proportions of species at each site are considered (as opposed to absolute abundances), the best index appears to be the Percentage Similarity Index, as it expresses the degree of community overlap on a linear scale and takes a wide range of values (Johannsson and Minns 1987). Hruby (1987) found the Bray-Curtis and Percentage Similarity Indices were decidedly superior to the nine other common indices he tested. It is frequently recommended that both these indices be based on logarithmically transformed data, to avoid bias by the most abundant species (Pontasch *et al.* 1989). Finally, in those rare instances where only presence-absence data are available, the Jaccard Coefficient is probably best (Hellawell 1977, Smith *et al.* 1990), although some maintain that the Coefficient of Community Loss (described in the next section) is superior (Courtemanch and Davies 1987, Plafkin et al. 1989).

It should be realized that no single similarity index will respond perfectly to every kind of difference between two communities, or at least not to every kind of difference that can theoretically exist. The Bray-Curtis Index and the Percentage Similarity Index have been shown in field experience to be the most reliable at tracking the kinds of changes observed between unperturbed communities in different places or between disturbed and undisturbed communities. These two indices have complementary biases; the Bray-Curtis is affected by absolute abundances while the Percentage Similarity Index considers only relative proportions. It may be useful, therefore, to plot both indices together, so that community changes of either kind will be included

Similarity indices appear to be a better choice for summarizing community structure than the alternatives discussed earlier, and they are useful tools for biomonitoring. The limitations of any summary statistic must be recognized, however, and similarity indices should serve along side of, and not in place of, more detailed examinations of individual taxa or functional groups. The sensitivity of similarity indices by themselves is limited by the natural variability among sites, and by the potential for small changes to be masked by the rest of the community. A dip in abundance of a few species at a downstream site, even if substantial compared with the controls, may cause only a slight change in community similarity between the sites because the abundances of all the unchanged species weigh in the calculation. Thus, similarity indices are best suited to summarize the degree of overall change between sites.

3.6.2 Coefficient of Community Loss

The inability of similarity indices to indicate the direction of change in community composition between two sites is seen by some as a major limitation of their use in pollution assessment. Most similarity indices are structured to show a decrease in similarity between an exposed and a reference site as the number of taxa in common decreases and as new species are recruited at the exposed site. Courtemanch and Davies (1987) suggest that a better index would consider the net loss of species at the exposed site, and proposed a new index, the Coefficient of Community Loss:

$$I = (a - c)/b$$

where:

a = Number of taxa in the reference community;

b = Number of taxa in the exposed community;

c = Number of taxa common to both communities.

The coefficient is the ratio of the number of taxa lost at the exposed site (a-c) to the number of taxa remaining (b), including any new taxa. The value of the index increases from zero, when no species are lost, to infinity, when no organisms remain at the exposed site. Recruitment of new taxa at the exposed site presumably conserves functional niches and essential community functions, so the equation is structured such that an increase in b reduces the numerical value of the coefficient.

The Coefficient of Community Loss is a re-working of an older index, the Species Deficit (Heckman *et al.* 1990), calculated as:

$$SD = (a-b)/a$$

The species deficit merely expresses the taxon richness of the exposed community as an (inverse) proportion of that of the reference community. It has a range from -1 to +1 and takes larger positive values as the exposed community loses species, negative values if it gains species. Unlike the Coefficient of Community Loss, however, the species deficit takes no account of the replacement of one species by another. Complete replacement of every species between the reference site and the exposed site would produce an *SD* of zero, i.e., no species loss.

Courtemanch and Davies (1987) tested the new index on a wide range of streams and rivers in Maine and claim that it was effective at differentiating sites suffering mild enrichment, at which species richness was increased relative to upstream controls, and those receiving severe organic loading or toxic effluents. Mild enrichment produced uniformly low index values, all <0.40, while severe impairment produced values above 0.80. The coefficient of community loss distinguished pristine and impaired sites more effectively than the commonly used Jaccard Coefficient.

The Coefficient of Community Loss is a step toward a similarity index that responds to the divergent responses of benthic invertebrate communities exposed to mild or severe disturbance. However, while in the test data of Courtemanch and Davies (1987) the index successfully separated different degrees of impairment (i.e., mildly enriched or severely polluted), it entirely failed to distinguish mildly enriched sites from pristine sites. In that respect the coefficient was less effective than a simple count of the number of species, which at least distinguished some enriched sites by their uniform increase in richness. Like all qualitative indices, the coefficient of community loss is very sensitive to rare species, and in the oligotrophic rivers of Maine this led to a wide variation in index values comparing two pristine sites. This new coefficient evidently has some value for classifying severely polluted sites, and it has been included as a metric in some rapid assessment methods (Resh and Jackson 1993). To date, however, an improvement in sensitivity to moderate disturbance compared with extant indices has not been demonstrated.

3.6.3 Per Cent Model Affinity

If upstream control sites are not possible, but data from a set of regional reference sites are available, Per Cent Model Affinity offers a method of classifying sites according to the degree of pollution impairment exhibited by the benthic invertebrate community. This method is still undergoing refinement, but at present it appears to be far too insensitive to be useful for routine biomonitoring at mines. As originally formulated (Novak and Bode 1992), Per Cent Model Affinity was a rapid assessment method for ranking river sites based on qualitative samples and gross taxonomic separations, intended to complement other water quality indices. The method establishes an expected or model community composition based on the reference sites, and then calculates affinity with that model using the Percentage Similarity Index. Communities with high similarity to the model are considered minimally disturbed, while lower similarities indicate increasing degrees of pollution.

The method proposed by Novak and Bode (1992) uses a qualitative kick-net sample from which the first 100 organisms are classified into one of only seven groups: Oligochaeta, Ephemeroptera, Plecoptera, Coleoptera, Trichoptera, Chironomidae, and Other. Using data from 23 pristine watercourses in New York State to define the model community, Novak and Bode showed that percentage similarity could be used to classify other sites in the state into four broad classes: unimpaired (similarity >65%), slightly impaired (50-64%), moderately impaired (35-49%) and severely impaired (<35%). The classifications agree with results from other indicators such as species richness and biotic indices, but this is hardly surprising given the range of severity of pollution effects being considered.

This first version of per cent model affinity was a coarse screening tool. While it could possibly have utility in regional water quality assessments, it would be worthless for biomonitoring at individual sites. Only the most dramatic changes in community composition would be detectable by this index, and for that purpose there is no shortage of established methods.

Barton (1996) modified the per cent model affinity method for application to streams draining agricultural land in southern Ontario, where pollution sources are diffuse and upstream controls are not possible because the entire drainage basin is often modified. Barton used qualitative samples (200 animals) but pursued taxonomy to the lowest practical level (genus or species for most groups) and improved the statistical rigour of site comparisons. The 69 reference sites were divided into subgroups based on season or habitat so as to produce groups with the lowest possible internal variability. The mean composition of these groups defined the reference communities for comparison with appropriate agricultural sites. Individual agricultural sites were judged to be significantly impaired if the per cent affinity with the reference community was less than the lower bound of the confidence limit for the mean of the model community. Barton (1996) found this approach to be effective for delineating environmental degradation from agriculture, especially when taxonomy was taken to the genus level.

Per cent model affinity differs from the use of similarity indices described earlier (Section 3.6.1) only in that the mean of regional reference sites is used as the basis for comparisons, instead of upstream control sites. There is no reason why any similarity index, not just percentage similarity, could not be used in this way. Barton (1996) and Novak and Bode (1992) point out that upstream controls will always produce superior results for analysis of individual sites along a river, and should be used whenever possible. Where that is not an option, regional reference sites may be the only answer.

3.7 Functional Feeding Groups

3.7.1 Theory and Practice

An alternative to analysis based on species or higher taxonomic groups is to classify species according to morphology or feeding habits into "functional groups" or guilds of similar organisms, regardless of taxonomic affinities (Cummins 1994). The idea of functional groups was originally envisaged as a way to circumvent taxonomic problems in benthos ecology and to provide a more meaningful classification of organisms according to their habits and habitats in the water course (Cummins 1973, 1974). Functional feeding groups are trophic guilds comprising macroinvertebrates which feed on the same food sources in the same manner, and thus have similar morphological and behavioural adaptations (Cummins 1994). The distribution of different functional groups varies according to the distribution of food resources (largely benthic algae, leaf litter and fine detritus), facilitating the understanding of organic matter processing in lotic waterways (Cummins 1974, Minshall *et al.* 1982). Proportions of different feeding groups have been shown to vary in predictable ways along a continuum from headwater streams to rivers (Hawkins and Sedell 1981, Cummins *et al.* 1981, Cushing *et al.* 1983).

Five feeding groups dominate most streams and rivers in North America: shredders, collectors-gatherers, collectors-filterers, scrapers and predators. Standard references now give functional group classifications for most insect taxa (Merritt and Cummins 1984), and others may be inferred from the literature. Recent attempts to extend the functional group concept to assessment of water quality is based on the reasoning that functional groups are a facet of the true ecological structure of the community. Therefore, stresses that disrupt the trophic dynamic of the system should be manifested in changes in functional feeding groups. Analysis of feeding groups would seem to offer a method of measuring impairment of ecological functions directly, rather than indirectly through comparisons of species distributions (Cummins 1994).

Pragmatically, functional groups also offer a defensible means of data reduction, because all the species at a site can be collapsed into five groups (or six, if a class for "other" is included). Analysis of functional groups measures real changes in the structure of the river ecosystem, rather than a convenient taxonomic or empirical index. While simplicity alone is a poor justification, if taxa must be lumped together for analysis, using feeding groups seems much more sensible than using arbitrary groups such as insect orders.

Functional feeding groups of benthic invertebrates reflect the food resources available in a given reach, so it follows that functional group distributions will respond most to disturbances that alter the food base of the system. For example, additions of light or nutrients increase algal production, which is followed by enhanced abundances of grazers (e.g., Hershey *et al.* 1988, Lamberti *et al.* 1989, Hart and Robinson 1990). Removal of riparian vegetation by logging or agriculture results in less leaf litter and a decline in populations of shredders (Dance and Hynes 1980, Tuchman and King 1993), but increases light penetration, which favours scrapers (Hawkins *et al.* 1982, Wallace and Gurtz 1986). Filterers tend to proliferate below lakes and reservoirs because of the export of suspended algae

(Richardson and Mackay 1991), but are reduced by inorganic suspended solids, which interfere with filtering (Taylor and Roff 1986).

These observations from stream ecology studies, coupled with field observations at polluted sites, suggest that functional groups can be expected to respond in different ways to different kinds of stresses: predators are often reduced in pollution-tolerant communities; sedimentation drastically reduces scrapers; while mild organic pollution tends to increase the proportion of filterers. The success of functional group analysis in ecological research suggests it holds considerable promise for biomonitoring. Cummins (1994) lays out a detailed sampling and analysis protocol for biomonitoring based on functional groups. One or more functional groups have now been included among the metrics for most rapid assessment methods (which attempt to rank water quality using a battery of easily measured attributes), including that used by the U.S. EPA (Barbour *et al.* 1992, 1996, Resh and Jackson 1993, Kerans and Karr 1994, Resh *et al.* 1995). Abundance of shredders in sub-alpine streams is a sensitive indicator of watershed acidification (Bruns *et al.* 1992).

The response of functional groups to pollution that does not obviously alter food resources is less clear. This is relevant to mines because toxic chemicals like heavy metals tend to eliminate species or reduce abundances but would not be expected to affect food resources appreciably. Still, heavy metal pollution of the Clinch River, Virginia, lead to proliferation of metal-tolerant blue-green algae and greater abundances of a metal-tolerant chironomid that fed on the algae (Clements *et al.* 1988a). Whether food resources would be altered by low-level, chronic metal loadings remains to be seen; Rasmussen and Lindegaard (1988) reported that the first species lost to chronic iron pollution in a Danish river were all scrapers, but that may have been coincidence, because they were all mayflies.

Even if functional feeding groups do respond in a predictable manner to disturbance, it is not yet established whether this method is sensitive to relatively small perturbations. Ecosystems tend to maintain processes at the expense of individual species (Schindler 1987); many species could be lost or replaced in a river reach without any drastic change in the distribution of feeding groups. Moreover, many invertebrates are opportunistic omnivores, changing their feeding habits to reflect the dominant food sources available. This large component of generalists may obscure responses to perturbation because the same species persist even after changes in the energy base of the stream.

Generalists are most common in flashy, unstable streams or those in boggy or northern areas where diversity is naturally low.

There is also the possibility that numerical abundances may not be the best way to express the distribution of functional feeding groups. The change in energy sources between forested and agricultural sites along a small stream in Michigan was obvious when group composition was expressed as biomass: the forested sites had a mixed community of shredders, collectors, grazers and predators, while the farmland sites without streamside vegetation were overwhelmingly dominated (>92%) by collectors-gatherers. But when expressed as population densities, no difference in community composition was evident between sites because the shredders and predators groups upstream were represented by low numbers of relatively large organisms. Hence in terms of abundance, all the sites appeared to be dominated by gatherers (Tuchman and King 1993).

Several experimental studies have purported to test the utility of functional feeding group analysis. Unfortunately, the value of most of these tests is limited because of inappropriate scales of measurement or lack of clarity about objectives. Those tests that have included clear comparisons of clean and disturbed sites have invariably used strong gradients of pollution, where the differences between communities would be evident by any method. Hence, the sensitivity of functional groups to moderate disturbance or subtle trends is still unknown.

Faith (1990) compared feeding group composition at 17 sites along the Upper La Trobe River system in Australia. They performed ordinations on abundances of each functional group and found no strong relationship among environmental variables except for stream order (a measure of size), substratum particle size and organic matter. More significantly, only the pattern for filterers, and perhaps scrapers, was stronger than that generated by random groups of invertebrate taxa without regard for feeding habits. However, this study encompassed a very long reach of river within which the anthropogenic influences are unspecified. The trends detected are those that would be expected along a river continuum; moreover, since Faith (1990) used absolute numbers in each functional group (as opposed to proportions) and total invertebrate densities increased along the length of the river, how could functional groups, or at least most of them, do otherwise?

The study by Palmer *et al.* (1996) on the Buffalo River, South Africa, at least included zones of known poor water quality. They studied feeding group distribution along the entire 100-km length of the river, which is disturbed by four reservoirs and agricultural, urban and industrial emissions. Their ordinations showed only a gradual change in functional group densities from headwaters to mouth; there was no gradient evident with respect to water quality. Again, however, conventional taxon-based analysis (ordination) also failed to detect any of the zones of poor water quality, and indeed reflected the same longitudinal trend evinced by feeding groups. At the large geographic scale of this study, the fundamental gradient of downstream change overwhelmed any local pollution effects in the ordinations, a shortcoming recognized by Palmer *et al.* (1996). More specific comparisons of sites above and below an effluent source are needed to properly test the value of functional groups.

Studies attempting to incorporate functional group analysis into more conventional environmental assessments have had greater success (Kondratieff *et al.* 1980, Olive *et al.* 1988, Camargo 1992b, Tuchman and King 1993). The general finding is illustrated in Table 3, which shows distribution of functional groups in a Maine river recovering from organic loads from a pulp mill (Rabeni *et al.* 1985). The cleanest sites support all groups, although the contribution of shredders will be low below the headwaters (Hawkins and Sedell 1981). Filterers tend to peak under mild enrichment while gatherers

Table 3. Distribution of invertebrate functional feeding groups among four classes of water quality (defined by polar ordination on species distributions) in a Maine river historically polluted with pulp mill effluent. (Source: Rabeni *et al.* 1985)

	Water Quality Class			
	I	II	III	IV
Functional Group	**Best**			**Poorest**
Shredder	2	<1	<1	<1
Scraper	45	37	4	<1
Gatherer	21	9	70	95
Filterer	22	39	70	1
Predator	9	15	18	4

dominate at severely polluted sites. For sites below a paper mill, Mayack and Waterhouse (1983) found good correlations between densities of functional groups and suspended solids, turbidity and rates of sediment deposition. They maintain that the feeding groups revealed changes in community structure that were not revealed by conventional analysis based on pollution tolerances. These examples illustrate the kind of information that can be extracted from functional group analysis, but they all concern organic pollution, and in every example the community changes were conspicuous and evident by other means.

Two published studies of benthic invertebrate feeding groups have been carried out at mines. Chadwick and Canton (1983) examined the effect of coal mine runoff on a stream in Colorado. Results showed an increase in filterers (the caddisflies *Hydropsyche* and *Cheumatopsyche*) below the first tailings piles and an increase in gatherers (*Ephemerella* and chironomids) at the farthest downstream site. Shredders and predators both declined sharply below the mine. Chadwick and Canton (1983) attribute the change to a shift in the energy basis of the stream from leaf litter to fine particulate organic matter, but it is uncertain whether runoff from the mine was responsible for the change.

Poulton *et al.* (1995) compared a number of simple metrics against metal concentrations in sediments and benthic invertebrates in the Clark Fork River, Montana, downstream from an active base metal mine. The distribution of functional groups (Figure 3) shows a progressive decrease in filterers and predators from the most contaminated site (CF1) through the two moderately contaminated sites (CF2 and CF3) to the three downstream sites. There were corresponding increases in gatherers from upstream to downstream, and perhaps surprisingly, a smaller but still considerable increase in shredders (Figure 3).

This example differs sharply from the earlier work on organic pollution in that filterers were most abundant at the most contaminated site, and gatherers only became dominant at less disturbed sites downstream. It is not yet known whether this effect is general.

3.7.2 Statistical Considerations

If functional feeding groups are to be used for biomonitoring, there is a need to formalize statistical procedures. Feeding groups have traditionally been compared among sites in any of three ways:

(1) as mean densities within each group;
(2) as proportions or percentages within each group;
(3) as ratios between two groups.

All three variables are commonly compared among sites using ANOVA, although in truth statistics are quite often neglected. Both of the first two methods require an ANOVA for each functional group. Popular ratios include scrapers/collectors (Poulton *et al.* 1995), scrapers/filterers and scrapers/(scrapers+filterers) (Barbour *et al.* 1992, Resh and Jackson 1993, Resh *et al.* 1995).

None of these expressions is completely satisfactory. The first expression includes both community structure, i.e., the proportions of different feeding groups, and population densities. Imagine an impaired site at which gatherers are reduced rather less than other groups. Gatherers could become the dominant group at that site, but the analysis would see only the decline in absolute abundance. The second expression may create a non-normal distribution, and there is a question of validity if all five functional groups are analyzed. When composition is expressed as proportions, the value of the last group analyzed is no longer an independent variable; it is determined by the magnitude of the other groups. Finally, ratios should be avoided where possible because they increase the variance compared with that of either variable alone, thus weakening the resolution of the index (Green 1979, Fore *et al.* 1996). There are 10 possible ratios among five functional groups, and the best ones may be different for different situations.

There is a fourth alternative for functional group analysis which deserves consideration. The real variable of interest is the distribution of invertebrates among feeding groups, which can be compared among sites with the Chi-square test (for counts) or the Kolmogorov-Smirnov two-sample test (for proportions). These methods do not require repeating an ANOVA calculation five times and they circumvent the technical issues of presentation discussed earlier. More importantly, a contingency table analysis directly addresses the real question, namely, whether the functional group distribution changed between sites, and it uses all the functional group data to make the determination. The potential for contingency table analysis of functional groups should be explored more thoroughly.

If functional analysis is to be construed as a complement to structural analysis of benthic communities, then perhaps it is valid to consider functional groups as analogous to taxa and analyze them accordingly. Both absolute numbers and proportions of taxa among sites are subject to comparisons, and each provides different information. The same may well apply to functional groups. Functional groups may be amenable to more complex methods such as MANOVA using two or more groups at once, or even to summary statistics like similarity indices. The last would provide a summary of functional similarity between two sites corresponding to the structural similarity indicated by conventional indices. These are ideas that are largely untried, and there is plentiful scope here for imaginative approaches.

The utility of functional feeding groups for biomonitoring at mine sites has not been established and begs further testing. Both theoretical and practical arguments can be made for the inclusion of functional groups, in that they are a key indicator of stream function, and are reliable enough for rapid assessment techniques. Questions remain, however, about whether functional groups are sufficiently sensitive to moderate disturbance, whether they contribute new information or are redundant with other measures, and whether their responses to disturbance are consistent enough to allow meaningful interpretation of biomonitoring data.

4. Statistical Methods

4.1 Basic Approach

The assessments considered in this review fall under the category of Green's (1979) *control-impact* comparison, where "impact is inferred from spatial pattern alone". Although it is primarily a semantic preference, we prefer to characterize the comparison as *reference versus exposed*, which implies fewer assumptions about both the unexposed and exposed areas sampled in the study than the terms *control* and *impact*. Similarly, the variable to be compared among sites, here called the *response variable*, is referred to as the *dependent* variable in some texts.

Although the simple spatial study design has important limitations (see Section 2.2), it is the design normally used in narrow-context assessments of the magnitude and spatial extent of a single outfall's effect on the receiving ecosystem. One of the important limitations of the design, potentially confounded causes for site differences other than effluent exposure, can be partially overcome by comparing sites with various levels of exposure to the mining effluent. In the typical study illustrated in Figure 1 there are three sites sampled and compared: reference (or control), high exposure and low exposure.

The statistical analysis of data from such a design is relatively straightforward. The null hypothesis is that there is no difference in the benthic community among sites. Whatever the response variable chosen to describe the community, the analysis will be some form of a one-way analysis of variance (ANOVA), sometimes referred to as a *single-classification* (Sokal and Rohlf 1995) or a *single-factor* (Zar 1996) analysis of variance. Virtually every statistical textbook, including Sokal and Rohlf (1995) and Zar (1996), provide the background, assumptions and calculations necessary to carry out a one-way ANOVA. Virtually every statistical software package, including the Statistical Analysis System (SAS), SPSS, and Minitab, include programs to carry out a one-way ANOVA. The necessary calculations can also be easily incorporated into computer spreadsheet programs.

The appropriate ANOVA table itself highlights the design of the study and should be presented for response variables with which hypotheses have been tested. If there are only two exposure levels (reference and exposed), data may be analyzed with a *t*-test (statistically equivalent to a one-way ANOVA with two groups), but it is probably better to use ANOVA so that covariates can be incorporated into the statistical model (see Section 4.3).

If more than one response variable is analyzed (e.g., three benthic species or five biotic indices), a multivariate analysis of variance (MANOVA) may be useful. Sometimes a MANOVA will reveal differences between sites not evident by looking at the response variables individually (Figure 4). The sample replication requirements for MANOVA are considerable, however. A reasonable rule of thumb is that at least five times as many observations as the number of response variables should be taken at each site. This requirement means 15 samples at both the reference and exposed sites would be needed to do a MANOVA with only three response variables. More observations are necessary in MANOVA to adequately estimate both the variances of each response variable and the pairwise covariance of response variables within each group of observations (Tabachnick and Fidell 1983).

If MANOVA is not possible, the potentially large number of ANOVAs to be performed on individual variables leads to an increase in the risk of falsely finding significant differences among sites, which is equivalent to the significance level (1-α) of the test being lower than the prescribed level. This problem is worsened when the variables to be analyzed are highly correlated, not an uncommon situation in benthic invertebrate data. For this reason the taxon list and derived variables should be carefully screened beforehand to avoid performing a large number of profitless analyses. A correlation matrix showing all the correlations among the variables selected for analysis can be quickly generated by any statistical software and will aid the researcher to discover which of the variables are redundant. Exhaustive testing of several variables may not be necessary if they are highly correlated and ANOVA on one reveals a significant difference among sites. Multivariate methods are also available to select or define uncorrelated variables from a matrix of taxon densities, but as discussed in Section 2.5, these methods themselves have limitations.

4.2 Alternatives to the Standard ANOVA

If the most important assumptions of ANOVA, normal distribution and equal variances of the response variable within each group, are not met, there are alternatives to the standard ANOVA hypothesis test. The most popular non-parametric version of the one-way ANOVA is the Kruskal-Wallis test. This analysis starts with a ranking of all response variable values, regardless of the site they came from, and ends with a statistical comparison of the ranks, rather than the original values among the groups. For example, it would test whether the reference site has more than its share, relative to the exposed site, of samples with the highest densities of an indicator species. Although this test can be almost as powerful as parametric ANOVA (Zar 1996), the lack of a quantitative, rather than just ordinal, difference among sites in the hypothesis test limits its interpretation.

Randomization tests, in which the data are randomly shuffled among the groups many thousands of times and the *F*-statistic recalculated, allow one to create a "home-made" distribution of the test statistic under the null hypothesis (Manly 1991). If the *F*-statistic calculated from the original, unshuffled data is unusually large relative to this home-made distribution, the null hypothesis is rejected. The significance of the rejection of the null hypothesis is just the proportion of the *F*-statistics in the home-made distribution that are greater than the *F*-statistic calculated from the original, unshuffled data. Manly (1991) provides Fortran programs for performing this kind of procedure, but with minimal facility with a standard statistical package like SAS, one could write a program to execute this sort of randomization test.

Another alternative to the standard ANOVA approach is more of a companion than an alternative, although it may be the only choice if the response variable is not amenable to statistical hypothesis testing. Clear graphical or tabular presentation of the data is critical in any report of comparisons among sites. Each figure or table should clearly show the pattern of difference among the sites in at least one response variable, and sometimes a covariate (see Section 4.3). It should also show the variability of the response variable within a site by showing either the scatter of the raw data, or standard error bars on a stem-and-leaf plot or box plot (Figure 5) for each site or exposure level.

4.3 Improving the Sensitivity of the Basic Approach

The sensitivity of the comparison of response variables between reference and exposed sites is often enhanced by either (1) constraining the sampling frame (e.g., sampling only riffle areas with flow between 50 and 75 cm/s) or (2) adding covariates to the comparison. Constraining the sampling frame may cause habitat-dependent changes in community structure to be missed if, for example, mine effluent affected only the benthic invertebrate community in pools. If any aspect of the habitat or environment, such as flow or substratum, is measured at each sampling point (as opposed to one measurement for each site), it may be useful to incorporate that habitat or environmental descriptor as a covariate in the statistical model (Figure 6). This makes the one-way ANOVA a one-way analysis of covariance (ANCOVA). Each quantitative covariate added to the model has one degree of freedom and n_e-1 degrees of freedom associated with the interaction between the covariate and exposure, where n_e is the number of exposure levels.

A complication is introduced into analysis of covariance if mean levels of the covariate itself differ between reference and control sites, for example, if current is faster downstream where effluent effects are greatest. This is a special case of confounding, since any difference in the communities inhabiting the sites may be at least partly because of the difference in the covariate (current velocity) and not the main variable (effluent concentration). In the worst case, adding the covariate to the model could eliminate a real difference between reference and exposed sites in the response variable. While it is still possible to conduct a covariance analysis when levels of the covariate differ among sites, the potential for a confounding error warns that the method should be applied with great caution. Perhaps repeating the analysis with and without the covariate would be the safest course. Confounding causes of site differences are a serious problem that should not be left up to ANCOVA to solve (see Section 4.4).

If there is an interaction between the covariate and the exposure factor in the ANCOVA, this may be interpreted as a significant effect of the effluent on the exposed community. It means that there is a difference between the reference and exposed sites in the relationship between the community and the covariate. For example, if mayfly density varied significantly with current velocity at the control sites but not at the exposed sites, it would suggest that the effluent was overwhelming the usual site preferences among these organisms, perhaps by exerting greater toxicity in fast-current areas. Depending on how the relationship differs between reference and exposed sites and the covariates and response variables involved, the interaction could be an ecologically important part of the effect of the effluent on the receiving system.

4.4 Limitations of the Basic Approach

Differences observed among the variously exposed sites could be caused by factors other than exposure to the effluent. Any differences among the sites in the relative frequency or the spatial and temporal variation of small-scale or large-scale habitat features may have caused the observed differences. Even a stochastic difference in the developmental rates of organisms at different sites could theoretically cause community differences that might be mistaken for effluent effects. This is the main criticism of the simple spatial design and related designs that rely on comparisons of only one site at each level of exposure to an effluent (Hurlbert 1984, Stewart-Oaten *et al.* 1986, Underwood 1991, 1992, 1994).

The primary method of avoiding, or at least evaluating, confounding is to collect habitat and environmental information from each site (such as degree of shading, slope, macrophyte cover, surrounding land use), so the potential for confounding can be qualitatively assessed. This is different from collecting information at the same scale as individual samples (e.g., current velocity, depth) for use as covariates (see Section 4.3). This site-scale information allows the investigator to see if differences in the benthic community among sites may have been suppressed or caused by variation among the sites in habitat or environmental conditions. Clearly, this qualitative analysis also requires independent information as to how the response variables used might be affected by conditions other than effluent exposure. Since variation in habitat conditions confounded with exposure variation may either (1) cause differences in the community that could be mistaken for effluent effects or (2) suppress differences in the community that will not be detected as effluent effects, care should be taken to evaluate the potential for confounding with all response variables, not just those for which a significant difference was detected. If more than one level of exposure has been sampled, then the chance of confounded cause is reduced if response variables show a pattern of variation corresponding to the level of exposure, i.e., most severe at the high exposure site, moderating at sites farther away from the effluent outfall (see Section 5).

In the special case of multiple exposure sources, it becomes particularly difficult to isolate one effluent as the reason for a change in the community. In this situation, if the reference site is not exposed to any of the effluents while the exposed site is exposed to several, it is clearly impossible to separate the various effects of individual effluent sources unless inferences are made from controlled studies (mesocosms) or from the proportion of effluent added by each source. If the reference site is exposed to all but the particular source of interest, the degree of difference in the exposed community may be either less than (if the community is already degraded) or greater than (if the effluent pushes the community's environment past a threshold) the difference that would appear if the reference site were pristine.

As in the case of natural confounding, if the spatial pattern of change in a response variable follows the intensity of a specific effluent exposure, the evidence of changes being caused by the effluent of interest is strengthened. Also, comparison of the reference community, which will be exposed to other effluents in this case, to other, similar, reference communities not exposed to disturbance may indicate the degree of degradation of the community prior to exposure to the particular effluent of interest.

As illustrated above, the one-way ANOVA model or its alternatives allow statistical decisions to be made on the magnitude of differences between reference and exposed sites relative to variability among observations within sites. The scale at which measurements are taken is important in the description and comparison of communities among sites. For example, consider a reference and control site that have each been sampled five times and the number of families of invertebrates in each sample tabulated, with the following results:

Reference Site	*Exposed Site*
13	9
8	14
15	11
12	14
12	10
mean = 12	mean = 12

The average number of families per sample is identical at the two sites, (12) but in fact there are twice as many families at the reference site than at the exposed site. The community varied more from sample to sample at the reference site, but this result is not discernable from the data presented

Differences between reference and exposed sites in, for example, the total number of taxa present, (as opposed to the average number of taxa per observation in each site) are not statistically testable. Nevertheless, this information could still be valuable as part of a weight-of-evidence decision as to the effect of the effluent on the community. Site-scale descriptors of the community should be calculated and presented in graphical or tabular form, and patterns of change in such variables should be interpreted as part of the study.

4.5 Ecological *versus* Statistical Significance

Section 2.4 defined power as the probability of rejecting the null hypothesis of no difference between the reference and the exposed community when there is indeed a difference. We only sample some of the organisms in the reference and exposed sites, so we never know what the true difference in the benthic invertebrate community is between them. If we decide on the amount of difference in a response variable that is important for us to detect (e.g., a decline in taxa richness of five per sample) and know the variability of the response variable and the sample replication used in the field, we can calculate power.

The mathematical calculation of power is straightforward, and formulae for power of *t*-tests and ANOVA are presented in many standard statistical texts (e.g., Zar 1996). Green (1989) discusses the details of power analysis for various ANOVA designs and a two-group MANOVA. When a minimum effect size for detection has been established, a power of 0.8, indicating an 80% chance of detecting site differences of the minimum size when such differences exist, is often taken as an acceptable limit for a good test (Peterman 1990a, Hyland *et al.* 1994). In practice, the exact magnitude of the acceptable power level will be a trade-off between the needs of the study and the available budget. Many routine biomonitoring studies muddle through with power levels that are depressingly low, because raising power to a respectable level is seen as prohibitively expensive. Routine use of power analysis would create an impetus for better funded studies and for using sampling programs and community metrics that have low variability and high power.

The real difficulty in evaluating power is determining what magnitude of difference in a response variable between reference and exposed sites is ecologically significant and therefore important to detect in a statistical comparison among sites. This is an issue of fundamental importance to biomonitoring which has engendered much debate. Choosing the minimum effect size for detection should not be a statistical decision based on the data gathered in a particular study. As Peterson (1993) has pointed out, there is a widespread fallacy in environmental assessment that an effect that is small in magnitude relative to natural variation is of no ecological consequence. It is the difference in the mean level of the variable that matters, and there are no adequate grounds for valuing a species less if it is more variable than another.

Nevertheless, the environmental effects monitoring program for pulp mills (DFO & Environment Canada 1995) suggests a difference in the means of the reference and exposed sites that is some multiple of the standard deviation of individual observations within either the reference or exposed sites. It further recommends the particular multiple 2σ, which implies that if the mean of the exposed site is outside of the 95% confidence interval of individual observations in the reference site, an important ecological difference exists.

This approach is potentially irrelevant to the real changes that occur as a result of exposure to mining effluent. A brief example illustrates this point. The number of invertebrate families found in a sample may be very constant from one point to another within a site (low σ), leading, with the above approach, to a powerful test of the difference in family richness between the reference and exposed sites. A small difference in richness per sample would be detected as significant with an unremarkable level of sampling effort.

Conversely, Hilsenhoff's Biotic Index may vary substantially from point to point within a site, meaning we have little power (relative to taxa richness) to detect a difference in biotic index between reference and exposed sites at the prescribed sample size. But what if the effluent of interest does not cause richness changes but does change the biotic index? We might be very good at detecting small changes in taxa richness with this sampling program, but that sensitivity is useless in our assessment context. To have substantial power to detect an ecologically important change in biotic index between reference and exposed sites, on the other hand, we may need to augment our level of replication.

The example above illustrates that minimum effect size to be detectable in a monitoring study should be an absolute number (e.g., a decline in species richness from 50 to 30 per sample) or a proportional amount (e.g., a 25% decline in filter-feeders), but not a multiple of variance at the reference sites (Green 1989). For regulatory purposes, acceptable power can be set by the responsible agency. A consensus decision might be reached by using expert opinion to nominate sites considered by all concerned to be clean reference sites or known degraded sites. This is not a circular procedure because the identification of known degraded sites is separate from the process of testing a site in a particular biomonitoring exercise. The difference between accepted "clean" sites and accepted "degraded" sites, or some fraction thereof, could be fixed as the minimum magnitude of difference that should be detected with some reasonable probability, where it is present. For example, if the difference between accepted clean and degraded sites was 50% in total abundance, then a difference of 20% might be decided as a threshold of degradation, and that change would be used to assess the power of a particular study to detect an ecologically important difference between reference and exposed sites.

There are limitations to this approach. The weight-of-evidence paradigm advocated here insists on measuring several or many variables. It would be daunting to establish minimum effect sizes for all of them in every region of the country. But setting minimum effect sizes for a few variables (total abundance and taxa richness are the most likely candidates) might discourage workers from using other variables.

It must be emphasized that the significant effect sizes under discussion here are *minimum* differences between sites that workers or regulators feel should be detectable by any reasonably well designed monitoring program. The minimum effect sizes are only for evaluating the statistical power of the test for differences among sites. A good study should strive to achieve *at least* the power to detect the minimum effect size. There is no prohibition implied against studies that have higher power than the suggested minima; indeed such high-power studies should be encouraged. This is not at all the same thing as a regulatory agency deciding minimum effect sizes to be considered ecologically or socially significant (e.g., if the change in species richness is <25% it will be neglected, for regulatory purposes, regardless of statistical significance), a proposition that never fails to arouse controversy. Power is adequate by definition in any test that finds a significant difference (Fairweather 1991), and it is up to the investigator to decide whether small but significant differences among sites are important ecologically or for environmental protection.

5. Interpretation of Statistics

5.1 Inferring Cause and Effect

As discussed earlier (Sections 2.2 and 4.4) biomonitoring at mines or other industries that are already in operation follows a variant of the simple spatial design, in which effects of the mine are inferred from differences in benthos community structure between unaffected control sites and a series or array of sites downstream or otherwise within the expected zone of influence (Environment Canada 1993). Inferential statistics with an explicit statement of power should be used to determine whether there are differences in community structure between upstream and downstream sites, and how far downstream those differences extend. What those differences imply for the affected community, however, is an ecological question which requires an understanding of the biology of various invertebrate groups and their places in the structure and function of aquatic ecosystems.

The major limitation of a spatial-design study is that differences between control and exposed sites can never be attributed to the intervention, in this case a mine site, with absolute confidence. It is always possible that the observed change in the benthic community would have arisen even in the absence of the mine, from some other external influence or simply from the natural variation from one place to another within the water body (Stewart-Oaten *et al.* 1986, Smith *et al.* 1993). Careful selection of sampling sites and coincident measurements of habitat variables at each site (water velocity, depth, substratum composition, etc.) will help to minimize the effect of habitat differences and other influences (Environment Canada 1993), but evidence of a mine effect remains circumstantial. Consequently, the case for an effect from the mine must be supported by additional evidence besides just significant differences in some aspects of the benthic community (Keough and Quinn 1991).

Support for a mine effect relies on three convergent lines of evidence:

(1) Observed differences at exposed sites consistent with expected effects of the mine, based on the nature of the habitat disturbance it creates or the materials in its effluent. To satisfy this premise requires a beforehand expectation of what kinds of influence the mine is going to exert, and what

sorts of effects to expect in the exposed community. While heavy metals are the most obvious contaminant to expect at a mine site, there is also the possibility of suspended or settling sediments from land disturbance or wastewaters, acidity from exposed acid-bearing rocks, organic wastes if domestic sewage is treated nearby, and nutrients from explosives or from fertilizers used in revegetation. Background information or effluent monitoring data are nearly always available to help formulate hypotheses of expected effects. Only those site differences that make sense with regard to the known properties of the mine effluent (etc.) can be used to support an argument of a mine effect.

(2) Reasonable consistency among the variables considered, sufficient to support a weight-of-evidence deduction that the mine effluent is responsible. For example, a decline in density or species richness of stoneflies and mayflies below a mine site might reasonably be taken as a response to a mine because at least some members of these orders are very sensitive to most sources of disturbance, including heavy metals. Other observations, however, must be congruent with the change in mayflies/stoneflies, if the supposition of a mine effect is to carry any weight. Smaller declines in other, more tolerant groups would strengthen the argument, but a significant increase in some species of mayfly or stonefly would cast doubt on the argument, unless there is evidence for competitive release. Similarly, if populations of filter-feeding caddisflies were depressed in the exposed reach, a decline in at least some species of mayflies and stoneflies would also be expected, because these latter orders are, by and large, more sensitive to metals (Clements *et al.* 1992, Kiffney and Clements 1994). On the other hand, if suspended sediment were the main effect, a decline in filter-feeders alone would not be inconsistent.

(3) A spatial pattern in abundance and community structure that is congruent with recovery of the community with distance away from the source, as the effluent effect attenuates. It is possible, if the stress is severe and long-lasting, that disturbance of benthic invertebrate communities will persist to the farthest sites with no evidence of recovery. In such rare situations, however, the effect of the mine would be obvious. The usual expectation, at least in the simplest case, is that the strongest effect would be just below the effluent or closest to the source of disturbance, and that populations would gradually recover farther away from the mine. Sequences of site differences that differ from

that pattern, especially those without any apparent spatial trend, should not be considered compelling evidence of a mine effect, irrespective of whether statistically significant differences are present.

There should also be at least a rough correlation between the magnitude of the effect on a given species, group or indicator and how far the disruption persists, relative to other metrics. Those taxa that are most severely affected by the effluent or disturbance would be expected to show the slowest rate of recovery. Taxa exhibiting smaller changes in abundance should show recovery at sites nearer to the source of disturbance (Clements *et al.* 1992).

5.2 Complications

All the discussion to this point has assumed the ideally simple case of a single point-source effluent discharging into a uniform water body. In the real world, where multiple discharges and disturbances are the norm, point-source discharges (tailings pond outflows) intermingle with diffuse sources (runoff from disturbed land or overburden), and tributaries, beaver dams, shifting currents and other vagaries of nature complicate field sites, consistent spatial patterns in benthic invertebrate populations may not be so readily discerned. Nevertheless, evidence of changes in the benthos at least consistent with a mine effluent effect is necessary if impairment is to be attributed to that source among several.

Confounding of effects is a persistent problem in biomonitoring (see Section 4). A number of investigators have recognized the difficulty of separating the effects of organic pollution and metals (LaPoint *et al.* 1984, Barton and Metcalfe-Smith 1992, Poulton *et al.* 1995, Peterson *et al.* 1996), and government agencies in the U.S. cite confounding of mixed-source pollution as one of the chief limitations of biomonitoring with benthic invertebrates (Van Hassel *et al.* 1988). How to attribute response of benthic organisms among several potential causes remains a major question to which more research should be devoted.

A further complication arises in the case of complex effluents, especially those that contain both toxins and nutrients. The toxic constituents in such effluents typically exert a pronounced effect on the benthic community immediately below the source. Sites farther afield will begin to show a stimulatory effect from the nutrients, once the effect of the toxins has been diminished through dilution or physical-chemical processes. The pattern among all exposed sites may then resemble an exaggerated, "over-recovery", in which abundances are greater at far downstream sites than at the control sites, but all species do not recover equally (Figure 7). A similar effect has been observed in the marine environment around oil rigs, which release both metals and hydrocarbons; at the same location near the rigs, some species are enhanced by the organic enrichment, while other are suppressed by the metals (Peterson *et al.* 1996).

An even greater challenge to interpretation is the possibility of a bi-directional response, an issue that has arisen throughout this review. Strong pollution invariably has clear disruptive effects on benthic communities and is relatively easy to confirm. Slight pollution, on the other hand, may produce either enhancement or suppression of individual population densities or some summary variables, especially diversity indices. The classic example is the response to mild enrichment, which increases food supplies and thereby allows new species to colonize without eliminating any of the original species (Novak and Bode 1992). Only when enrichment becomes more severe do species replacements and losses begin.

Similar stimulation responses are not unknown in the presence of other sorts of disturbance, including sedimentation and metals. For example, increases in richness of Ephemeroptera and Plecoptera species have been noted in some streams after clearcutting of the surrounding forest because the disturbance allows new species to colonize from downstream but is not severe enough to eliminate any of the original species (Wallace *et al.* 1996). Perhaps surprisingly, moderate levels of heavy metal pollution may also be stimulatory, producing populations of chironomids and filter-feeding caddisflies (family Hydropsychidae) substantially greater than those at sites without contamination (Clements *et al.* 1988a). The possibility of a bi-directional response complicates analysis because either an increase or a decrease in the variable of interest could be indicative of slight impairment.

A change in the benthic invertebrate community in the opposite direction from that expected is a special case of a more general issue: what is to be concluded from significant, substantial, internally consistent changes in the exposed zone that were not predicted beforehand? There is an element of circularity in supposing an effect is attributable to the mine after the fact. On the other hand, given the complexity of benthic assemblages it is always possible, in fact likely, that some of the changes in the exposed community will be unanticipated, either because of our lack of understanding of the ecosystem or because the composition of the effluent or the nature of the disturbance was different than what was supposed.

Properly, the only objective statement that can be made is that an unexpected trend occurs downstream from the mine and that no other plausible source can be found. It is possible to treat the unexpected effect as a hypothesis, and look for other effects that are consistent with it. For example, if stimulation of one species were detected downstream where toxicity had been expected (that is, population densities increased instead of declining), an *ad hoc* hypothesis could be formed that there is stimulation from an unknown source and a search undertaken for other evidence of the same effect. Confirmatory findings would support the existence of stimulation, but to ascribe that finding to the mine would require new or additional information about the nature of the effluent. It would be circular to attribute the unexpected effect to the mine in the absence of outside evidence. The scientific literature is a rich source of information on species sensitivities and responses to pollutants of different kinds, and can provide much of the independent information needed to confirm a probable cause-effect link between mine effluents and an unanticipated response in the benthic community.

5.3 Effects of Mines

It is beyond the scope of this report to annotate all the possible responses of benthic invertebrates to disturbances of different kinds. As a broad generalization, there is good evidence now that stoneflies, mayflies and cased caddisflies, in that order, are the most sensitive groups in most benthic assemblages, and that chironomids and oligochaetes are more tolerant (Winner *et al.* 1980, Klemm *et al.* 1990). The tendency of those first three insect orders to respond to relatively small disturbances is the basis for the EPT index, which does work well for many kinds of disturbance (Lenat 1988, Quinn and Hickey 1990, Kerans and Karr 1994), including metals (Poulton *et al.* 1995). Hence a decline in abundances or species richness among the EPT taxa is often the first indicator that biologists look at when analysing results of a benthic survey.

However, on closer examination, generalizations like the EPT index are riddled with exceptions. While stoneflies, especially stenothermal leaf-shredding species, are often considered the most

sensitive taxa to organic enrichment, they are generally less sensitive than mayflies to heavy metals (Table 4). A number of studies have now shown that stoneflies and caddisflies are quite tolerant of metals, and among the first species to recover at metal-contaminated sites (Clements *et al.* 1988b, 1992 and references therein). Hydropsychidae are sometimes the dominant species at sites of moderate metal contamination (Clements *et al.* 1988a, Poulton *et al.* 1995).

The responsiveness of the EPT orders, as well as the purported resistance of the others, is a consequence of the relative sensitivities of some of their member species. Each of these groups contains tolerant and intolerant members to different kinds of stress; the apparent sensitivity or tolerance of the group arises from the presence of some species that stand out for being very tolerant or very sensitive. This is especially true for metals, in which sensitivity appears to be extremely species-specific (Figure 8), and varies even from one metal to the next.

As an example, Table 4 presents relative sensitivities of immatures of various insect species to heavy metals of the nature and concentration found at mine sites in the U.S. At the level of the order, the EPT taxa were all decidedly more sensitive to copper or a three-metal mixture than chironomids, some or all of which showed improved survival relative to control streams, while the other species declined. Within each order, however, there is a wide range of sensitivities for individual species or higher taxa. The higher concentration of copper eliminated *Isonychia bicolor* and *Pseudocloeon* sp., but reduced *Tricorythodes* sp. by about two-thirds. Within the Chironomidae, it is only the tribe Orthocladiini that increased in abundance in copper-exposed streams, while a lower dose of three metals stimulated all the subfamilies. Many studies have found that orthoclad midges thrive, or at least survive, under quite severe metal stress, whereas tanytarsids are as sensitive as mayflies (Winner *et al.* 1980, LaPoint *et al.* 1984, Leland *et al.* 1989).

Tolerance to heavy metals evidently varies widely even within genera, as is strikingly demonstrated by the large mayfly genus *Baetis*. Clements *et al.* (1992) found *Baetis brunneicolor* among the most sensitive insect species to copper, and Kiffney and Clements (1994) report similarly high sensitivity of *B. tricaudatus* to a mixture of metals (Table 4). Conversely, Leland *et al.* (1989) found that unspecified species of *Baetis* were the second most tolerant among five genera exposed to copper in laboratory streams (Figure 8). Rasmussen and Lindegaard (1988) sampled benthos from Danish streams containing a very wide range of dissolved iron concentrations; some species of *Baetis* disappeared in streams with iron concentrations as low as 0.2 mg/L, but at least one species was found in all but the most heavily contaminated streams (10 mg/L). The wide range of metal tolerances within this genus has lead to contradictory conclusions: Clements *et al.* (1988a) concluded that *Baetis* was quickly eliminated from the mine-contaminated Clinch River because of its high sensitivity to metals; Roline (1988), on the other hand, found *Baetis* at most metal-contaminated sites on another river system and concluded that it was a relatively tolerant genus.

Again, most of the examples cited here that have examined effects of metals on species richness within specific groups of invertebrates have been based upon severely contaminated sites or experiments employing substantial acute doses of heavy metals. The extension of these results to less drastic metals loads is not obvious. Responses of the benthos to chronic, low-level metals loads are likely to be expressed as relatively simple changes, i.e., loss of a few very sensitive species and reductions in abundances of others. Erection of hypotheses concerning expected effects may be simpler under this scenario than in the more complex situations discussed earlier. However, much more guidance on the effects of low-level metals contamination, along with other effects of mining, are needed if biomonitoring is to be optimized.

Table 4. Change in relative abundances of various taxa of benthic insects in outdoor artificial streams dosed with copper (25 µg/L) or copper (12 µg/L), cadmium (1.1 µg/L) and zinc (110 µg/L) for ten days. (Sources: Clements *et al.* 1992, Kiffney and Clements 1994)

Taxon	Copper	Copper Cadmium Zinc	Taxon	Copper	Copper Cadmium Zinc
Ephemeroptera		-0.68	Trichoptera		
Isonychia bicolor	-1.00		*Neureclipsis* sp.	-0.53	
Pseudocloeon sp.	-1.00		*Cheumatopsyche* sp.	-0.30	
Baetis brunneicolor	-0.99		*Hydropsyche bifida*	-0.16	
B. tricaudatus		-0.76	*Lepidostoma ormeum*		-0.17
Caenis sp.	-0.95		Chironomidae		0.56
Stenonema modestum	-0.73		Tanytarsini	-0.79	0.84
Tricorythodes sp.	-0.67		Chironomini	-0.70	0.80
Drunella grandis		-0.65	Tanypodini*	-0.30	0.94
D. doddsi		-0.66	Orthocladiini*	0.14	0.15
Heptageniidae		-0.90			
Plecoptera		-0.44			
Pteronarcella badia		-0.60			
Suwallia pallidula		-0.36			
Sweltsa coloradensis		0			

* Tribe in first column (Tanypodini, Orthocladiini), subfamily in second column (Tanypodinae, Orthocladiinae).

5.4 Benthic Invertebrates and Stream Processes

The point of biological monitoring for environmental protection is to measure the effects of waste discharges or land disturbance directly, through the responses of indigenous organisms, rather than extrapolating from chemical and physical measurements. Benthic invertebrates have been favoured for this purpose for a number of often-cited reasons: they are sensitive to changes in both the physical and chemical quality of their environment, are reasonably easy to collect and identify, and are present in large numbers on or near the substratum at all times of the year. Moreover, benthic invertebrates occupy an intermediate position in the trophic structure of aquatic ecosystems, between primary producers (algae and aquatic plants) on the one hand, and large predators (mostly fish and waterfowl) on the other. As intermediate consumers, macroinvertebrates are influenced by both competition resulting from food limitations (termed bottom-up forces in the ecological literature) and predation from larger predators (top-down forces) and serve as conduits through which these effects are transmitted to other trophic levels (Wallace and Webster 1996).

Macroinvertebrates can have important effects on nutrient cycles, primary production, decomposition and translocation of materials. Given their central role in energy flow and organic matter transformation, disruption of invertebrate communities presumably has ramifications for the entire stream or lake ecosystem. Environmental protection efforts that incorporate benthic invertebrate monitoring are concerned with more than just the protection of the invertebrate population itself; impairment of these communities also serves as a warning of threats to other biota, including fish, and of degradation of the ecosystem as a whole (Reice and Wohlenberg 1993).

The integrity of an aquatic ecosystem, whether lotic or lentic, depends upon the maintenance of essential ecosystem functions such as primary and secondary production, element cycling (or spiralling, in streams), carbon transformations and decomposition. While measurements of benthic invertebrate abundance and community structure are relatively straightforward, unravelling the complex linkages between benthos structure and ecosystem processes remains a challenge for ecologists (see Reice and Wohlenberg 1993 for a review). Wallace and Webster (1996) concluded from a recent review that the extent to which classic environmental quality indicators like biotic indices and modified invertebrate community structure indicate altered ecosystem-level processes for many kinds of anthropogenic disturbances remains unknown.

Probably the best example of the importance of benthic invertebrates to ecosystem-level functions in flowing waters, at least, comes from a multi-year study of a headwater stream in the Appalachian Mountains (North Carolina) repeatedly dosed with insecticide (Cuffney and Wallace 1989, Cuffney *et al.* 1990, Wallace *et al.* 1991, Lugthart and Wallace 1992). The study stream was treated with a heavy dose of methoxychlor, an insecticide frequently used to control blackflies, and treatment was repeated two or three times per season for two years. The insecticide caused massive catastrophic drift and drastically reduced the population density of benthic invertebrates. Shredders and filterers were practically eliminated, and the remaining functional groups were dominated by Diptera and non-insect taxa such as oligochaetes (Cuffney and Wallace 1989). Annual production by all functional feeding groups except gatherers, the dominant group during treatment, declined by 71-94% compared with pre-treatment levels (Lugthart and Wallace 1992).

The insecticide treatment had equally pronounced effects on energy flow and detritus processing. Compared with paired-basin control streams, leaf litter processing by shredders was depressed 50-74%, and export of fine particulate matter was reduced by a third (Cuffney *et al.* 1990). The magnitude of fine particulate matter export during storms and the seasonal pattern of export were also profoundly altered by the reduction in invertebrate populations (Wallace *et al.* 1991). The insecticide treatment had a three-times greater effect on fine particulate matter transport than a 50-year drought (Cuffney and Wallace 1989).

Response of the benthic invertebrate community to the insecticide treatment was measured in the treated stream and a nearby control stream throughout the three-year disturbance and the first two years of recovery (Wallace *et al.* 1996). Biomonitoring used two simple measurements: EPT taxon richness and the North Carolina Biotic Index, a quantitative index based on tolerance values weighted by population density (Lenat 1993). The effect of methoxychlor on EPT taxa was immediate and devastating. Richness declined from 13-21 taxa before treatment to 2-8 taxa, and remained at that level for the duration of the treatment. A full year of recovery was required for the lost taxa to recolonize. The biotic index similarly reported a sharp increase at the onset of insecticide treatment which only returned to pre-treatment levels after the treatments ended.

There was a very close correspondence between reductions in EPT taxon richness or increases in the biotic index and changes in organic matter transport and processing (Figure 9). Decomposition rates of leaf litter during insecticide treatment were roughly half the rate before or after treatment, and the depression and recovery of decomposition rates corresponds very closely with EPT taxon richness (Figure 9A), reflecting the influence of shredders. Similarly, there is a near-perfect inverse correspondence between EPT taxon richness and fine particulate transport (Figure 9B), a consequence of the previously discussed role of invertebrates in shredding large detritus (Wallace *et al.* 1996).

This study confirms the linkage between invertebrate community structure and ecosystem function in streams, and indicates that conventional biomonitoring measurements do have ecological meaning. However, the stress imposed by insecticide treatment in this stream was catastrophic, and only the most severe instances of contamination from industry would be equivalent. Effects of less extreme disturbances would presumably be commensurate with the degree of change in the benthic community, but field research to support this supposition remains wanting.

Nevertheless, if the goal of biomonitoring is to foresee and prevent degradation of aquatic ecosystems, there remain compelling reasons for using invertebrate abundances and community structure as its foundation. First among them is the far greater sensitivity of species and populations to perturbation compared with ecosystem functions. Based on many years of research in the Experimental Lakes Area of Ontario, Schindler (1987) pointed out that lake ecosystem processes such as primary production, decomposition and phosphorus cycling are largely unaffected by stresses like heavy metals or acidification unless the stress is severe or persists for a long time. Individual species, on the other hand, especially fast-dispersing organisms with short life cycles, are reliable indicators of pollution effects.

This conclusion makes sense given the natural redundancy of invertebrate communities (Wallace and Webster 1996). If a single sensitive species is eliminated from a stream riffle, the unused food resource will still be exploited, either by new colonizing species, or more commonly, by a slight expansion of feeding habits by other species already present. The net effect on energy and materials flow is negligible. It is only when the redundancy among species is exhausted by severe perturbation, as in the study of Wallace *et al.* (1996), that effects on processes become evident.

Coupled with this are more pragmatic arguments based on the relative ease and speed of collecting and counting organisms compared with measurements of production or other processes, and the greater information content of a benthic invertebrate community. On the other hand, Lugthart and Wallace (1992) contend that abundance greatly underestimates the severity of disturbance compared with production, at least for the severe disturbance they studied. Steep declines in abundance therefore may herald even greater disruption of fundamental processes.

6. Conclusions and Recommendations

6.1 Conclusions

This review has been built upon the premise that a weight-of-evidence approach based on inferential statistics and graphical presentation of data is the most effective means of analysing benthic invertebrate counts from a biomonitoring study. Hypothesis-testing statistics, univariate and multivariate, are the most effective tools to establish differences among sites, but must be coupled with an understanding of the biology of the aquatic ecosystem potentially affected by the mine and the nature of the disturbances to which it has been subjected. Descriptive, multivariate statistics are an important adjunct to biomonitoring but are not preferred as a means of establishing whether invertebrate communities below effluent outfalls or other sources of disturbance associated with mining have been significantly degraded compared with reference communities. Strong inference associating the mine with observed biological effects requires independent information about the supposed effects of the effluent (or other disturbance), and a careful examination of the spatial pattern of degradation and recovery among exposed sites.

Given this context, three main conclusions emerge from the review.

(1) The determination of effects of mines is strongest when it is based on the composite results for many taxa and community variables rather than a few summary statistics. Individual species or higher taxa are the most varied and sensitive indicators of environmental conditions. The parallel analysis of several taxa provides both an opportunity to confirm the direction of observed trends and invaluable insights into the nature of the stresses affecting the community. Nevertheless, selected summary statistics ought to be included in the analysis to provide a measure of the severity of effects on the community as a whole.

(2) Most methods already in routine use and most new methods advocated in the scientific literature have not been tested adequately for their utility at detecting subtle effects or making relatively fine discriminations between pristine and mildly impaired sites. Sensitivity is necessary if biomonitoring is to function as an early warning system of incipient degradation, rather than a means of describing catastrophe. Yet the predominance of field assessments of methods have used seriously perturbed sites at which the disruption of benthic communities at the exposed sites could hardly fail to be noticed by any means. Careful tests of measures and methods at less severely disturbed sites are sorely needed.

(3) Statistical power is a key element of sound experimental design in biomonitoring that has not been afforded the attention that it deserves. A growing number of researchers and practising biologists do recognize the importance of power in field studies and are incorporating power analysis into their programs (e.g., Hyland *et al.* 1994). But the full benefits of power analysis in improved sensitivity and efficiency of biomonitoring have yet to be realized. Power analysis should be routinely incorporated into every biomonitoring study, both in the formulation of sampling plans and in the assessment of nonsignificant variables during data analysis.

6.2 Recommendations

(1) Analysis of Variance or its derivatives (ANCOVA, MANOVA) is the preferred method of testing for significant differences in species abundances or community metrics among sites in a biomonitoring study. Analysis of Covariance is a powerful means of reducing variability and thereby increasing the sensitivity of the analysis, and its careful use is to be encouraged. The use of

Multivariate Analysis of Variance, otherwise a very promising tool, is restricted in routine monitoring studies by the requirement for large numbers of replicates. Modifications to sampling programs (collection of habitat data at each sampling point, increasing replication and decreasing sample size) that would facilitate the use of these two methods should be promoted.

(2) Abundances of common taxa, aggregated if necessary, should be the keystone of the analysis for site differences. Higher taxonomic levels such as insect orders should only be used where lower taxa are too rare or too variable to be useful and the members of the higher taxon are reasonably similar in ecological requirements. Total abundance of all organisms and total number of taxa per sample are useful variables but may be unresponsive to slight degradation. A similarity index, or possibly two, should be included in the analysis as a means of expressing the net change in community structure between sites.

(3) Diversity indices tend to be unresponsive to slight or moderate disturbance, especially when it does not involve organic enrichment, and are not recommended for biomonitoring at Canadian mine sites. Biotic indices may be useful, but should only be included as part of the analysis of benthic data from mines when they are applicable to the geographic region and there is reason to expect organic or mixed effluents. Biotic indices must be calculated for each sample and subjected to statistical analysis in the same manner as other variables. All metrics based on ratios between two variables should be avoided.

(4) The utility of functional feeding groups as variables to estimate impairment of benthic communities at mine sites is uncertain. The theoretical basis for including functional measures in an assessment is attractive, but research to date has laid too much emphasis on evaluating effects of severe impairment. More research is needed to test the sensitivity and reliability of feeding groups at moderately contaminated sites where the food base has not been directly altered.

(5) Some metrics used in rapid assessment procedures may also be useful in quantitative biomonitoring, and research comparing the sensitivity and accuracy of different metrics should not be disregarded. However, the "multi-metric" approach to biomonitoring, in which a diversity of

unrelated metrics are combined into a single number to rank sites, is not sound biologically or statistically and is to be strongly discouraged.

(6) The use of simple graphs of species abundances, richness or other variables against sites and distances from point sources should be encouraged as a straightforward and easily comprehended means of presenting benthic invertebrate data. Means and ranges or standard deviations should be included on the graphs along with an indication of statistically significant differences. Graphical depictions of site descriptors (e.g., canopy cover, land use, discharge, slope) that cannot be statistically compared are also useful as a means of exploring and illustrating large-scale differences among sites. Graphs from ordinations or clustering dendrograms can also be informative but should not displace simple scatterplots of the original data as the mainstay of data presentation.

(7) Statistical power calculations should become a standard component of every biomonitoring study, and should be included in the report. Power should be calculated during study design, based on preliminary sampling or data from previous years, to ensure that sampling intensity is sufficient to ensure a reasonable probability of detecting site differences of a magnitude deemed to be ecologically significant. Power analysis should be applied during data analysis to every analysis of variance that fails to detect a significant difference among sites. The power analysis should either demonstrate that the power of the test was reasonable, or determine the magnitude of difference between sites that would be required for a test of reasonable power. For tests that have low power, determining the sampling intensity that would be necessary to ensure a powerful test would aid in planning of subsequent studies. Persistently low power may even indicate that a different monitoring tool besides benthic invertebrates should be considered.

(8) More research on the effects of mine wastes on benthic invertebrates in lakes and rivers, especially their responses to low-level, chronic loading and to mixed metal-organic wastes, would help investigators attempting to formulate hypotheses of expected mine effects. Research to determine the occurrence and significance of stimulation responses at slightly contaminated sites is especially recommended because of the complexity of interpretation introduced by bi-directional responses to disturbance. Laboratory and mesocosm experiments to establish toxicity of various metals to a variety of common benthic species in Canadian water bodies would also be welcomed.

(9) All raw data from each biomonitoring study should be archived in a safe, organized, accessible data base for future studies of temporal trends, and possible integration into a network of regional reference sites.

7. Literature Cited

AFNOR (Association Française de Normalisation). 1985. Essais des eaux. Détermination de l'indice biologique global (IBG). AFNOR T90-350, October 1995. 8 p. (Cited in Metcalfe 1989).

Agard, J.B.R., Gobin, J. and Warwick, R.M. 1993. Analysis of marine macrobenthic community structure in relation to pollution, natural oil seepage and seasonal disturbance in a tropical environment (Trinidad, West Indies). Mar. Ecol. Prog. Series. 92: 233-243.

Anderson, A.M. 1990. Selected methods for the monitoring of benthic invertebrates in Alberta rivers. Environmental Quality Monitoring Branch, Environmental Assessment Division, Alberta Environment, Edmonton, Alberta. 41 p.

Armitage, P.D. and Blackburn, J.H. 1985. Chironomidae in a Pennine stream system receiving mine drainage and organic enrichment. Hydrobiologia 121: 165-172.

Armitage, P.D., Moss, D., Wright, J.F. and Furse, M.T. 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Res. 17: 333-347.

Barbour, M.T., Gerritsen, J., Griffith, G.E., Frydenborg, R., McCarron, E., White, J.S. and Bastian, M.L. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. J. N. Amer. Benthol. Soc. 15(2): 185-211.

Barbour, M.T., Plafkin, J.L., Bradley, B.P., Graves, C.G. and Wisseman, R.W. 1992. Evaluation of EPA's rapid bioassessment benthic metrics: Metric redundancy and variability among reference stream sites. Envir. Toxicol. Chem. 11: 437-499.

Barton, D.R. 1996. The use of Percent Model Affinity to assess the effects of agriculture on benthic invertebrate communities in headwater streams of southern Ontario, Canada. Freshwat. Biol. 36: 397-410.

Barton, D.R. and Metcalfe-Smith, J.L. 1992. A comparison of sampling techniques and summary indices for assessment of water quality in the Yamaska River, Québec, based on benthic macroinvertebrates. Environ. Monitor. Assess. 21: 225-244.

Battegazzore, M., Petersen, R.C., Moretti, G. and Rossaro, B. 1992. An evaluation of the environmental quality of the River Po using benthic macroinvertebrates. Arch. Hydrobiol. 125: 175-206.

Beak Consultants Ltd. 1990. Technical guidance manual for aquatic environmental effects monitoring at pulp and paper mills. Vol. 2. Procedures. Chapter 10: Benthic community assessment. 12 p.

Beisel, J.-N., Thomas, S., Usseglio-Polatera, P. and Moreteau, J.-C. 1996. Assessing changes in community structure by dominance indices: A comparative analysis. J. Freshwat. Ecol. 11(3): 291-300.

Bernstein, B.B. and Zalinski, J. 1983. An optimum sampling design and power tests for environmental biologists. J. Envir. Manage. 16: 35-43.

Bloom, S.A. 1981. Similarity indices in community studies: Potential pitfalls. Mar. Ecol. Prog. Ser. 5: 125-128.

Boyle, T.P., Smilie, G.M., Anderson, J.C. and Beeson, D.R. 1990. Sensitivity analyses of nine diversity and seven similarity indices. J. Water Pollut. Control Fed. 62: 749-762.

Bray, J.R. and Curtis, J.T. 1957. Ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27: 325-349.

Brinkhurst, R.O. 1993. Future directions in freshwater biomonitoring using benthic macroinvertebrates. *In:* Rosenberg, D.M. and Resh, V.H. (editors) 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, NY. p. 442-460.

Bruns, D.A., Wiersma, G.B. and Minshall, G.W. 1992. Evaluation of community and ecosystem monitoring parameters at a high-elevation, Rocky Mountain study site. Environ. Toxicol. Chem. 11: 459-472.

Brussock, P. 1993. Experimental design and controls: Key to refining rapid bioassessment metrics. Envir. Toxicol. Chem. 12: 1-4.

Bunn, S.E. 1995. Biological monitoring of water quality in Australia: Workshop summary and future directions. Aust. J. Ecol. 20: 220-227.

Camacho, F. and Vascotto, G.L. 1991. Framework for enhancing the statistical design of aquatic environmental studies. Environ. Monitor. Assess. 17: 303-314.

Camargo, J.A. 1992a. New diversity index for assessing structural alterations in aquatic communities. Bull. Environ. Contam. Toxicol. 48: 428-434.

Camargo, J.A. 1992b. Structural and trophic alterations in macrobenthic communities downstream from a fish farm effluent. Hydrobiologia 242: 41-49.

Chadwick, J.W. and Canton, S.P. 1983. Coal mine drainage effects on a lotic ecosystem in Northwest Colorado, U.S.A. Hydrobiologia 107: 25-34.

Chadwick, J.W. and Canton, S.P. 1984. Inadequacy of diversity indices in discerning metal mine drainage effects on a stream invertebrate community. Water, Air, Soil Pollut. 22: 217-223.

Chutter, F.M. 1972. An empirical biotic index of the quality of water in South African streams and rivers. Water Research 6: 19-30.

Clements, W.H., Cherry, D.S. and Cairns, J. Jr. 1988a. Impact of heavy metals on insect communities in streams: A comparison of observational and experimental results. Can. J. Fish. Aquat. Sci. 45: 2017-2025.

Clements, W.H., Cherry, D.S. and Cairns, J. Jr. 1988b. Structural alterations in aquatic insect communities exposed to copper in laboratory streams. Environ. Toxicol. Chem. 7: 715-722.

Clements, W.H., Cherry, D.H. and Van Hassel, J.H. 1992. Assessment of the impact of heavy metals on benthic communities at the Clinch River (Virginia): Evaluation of an index of community sensitivity. Can. J. Fish. Aquat. Sci. 49: 1686-1694.

Cook, S.E.K. 1976. Quest for an index of community structure sensitive to water pollution. Environ. Pollut. 11: 269-288.

Courtemanch, D.L. and Davies, S.P. 1987. A coefficient of community loss to assess detrimental change in aquatic communities. Water Res. 21: 217-222.

Crunkilton, R.L. and Duchrow, R.M. 1991. Use of stream order and biological indices to assess water quality in the Osage and Black river basins of Missouri. Hydrobiologia 223: 155-166.

Cuff, W. and Coleman, N. 1979. Optimal survey design: Lessons from a stratified random sample of macrobenthos. J. Fish. Res. Board. Can. 36: 351-361.

Cuffney, T.F. and Wallace, J.B. 1989. Discharge-export relationships in headwater streams: The influence of invertebrate manipulations and drought. J. N. Amer. Benthol. Soc. 8: 331-341.

Cuffney, T.F., Wallace, J.B. and Lugthart, G.J. 1990. Experimental evidence quantifying the role of benthic invertebrates in organic matter dynamics of headwater streams. Freshwat. Biol. 23: 281-299.

Cummins, K.W. 1973. Trophic relations of aquatic insects. Ann. Rev. Entomol. 18: 183-206.

Cummins, K.W. 1974. Structure and function of stream ecosystems. BioScience 24: 631-641.

Cummins, K.W. 1994. Bioassessment and analysis of functional organization of running water ecosystems. *In:* Loeb, S.L. and Spacie, A. (editors). 1994. Biological monitoring of aquatic systems. Lewis Pub., Boca Raton, FL. p. 155-169.

Cummins, K.W., Klug, M.J., Ward, G.M., Spengler, G.L., Speaker, R.W., Ovink, R.W., Mahan, D.C. and Peterson, R.C. 1981. Trends in particulate organic matter fluxes, community processes and macroinvertebrate functional groups along a Great Lakes Drainage Basin continuum. Verh. Internat. Verein. Limnol. 21: 841-849.

Cushing, C.E., McIntire, C.D., Cummins, K.W., Minshall, G.W., Petersen, R.C., Sedell, J.R. and Vannote, R.L. 1983. Relationships among chemical, physical, and biological indices along river continua based on multivariate analyses. Arch. Hydrobiol. 98: 317-326.

Dance, K.W. and Hynes, H.B.N. 1980. Some effects of agricultural land use on stream insect communities. Environ. Pollut. (Series A) 22: 19-28.

De Pauw, N. and Vanhooren, G. 1983. Method for biological quality assessment of watercourses in Belgium. Hydrobiologia 100: 153-168.

DFO (Department of Fisheries and Oceans) and Environment Canada. 1995. Further guidance for the invertebrate community survey for aquatic environmental effects monitoring related to the federal *Fisheries Act* requirements. Environmental Effects Monitoring Program, Environmental Conservation Service, Environment Canada, Ottawa, Ontario. Report EEM 2. 206 p.

Environment Canada. 1993. Guidelines for monitoring benthos in freshwater environments. Prepared for Environment Canada, Pacific and Yukon Region, North Vancouver, B.C. by EVS Consultants, North Vancouver, B.C. 81 p.

Fairweather, P.G. 1991. Statistical power and design requirements for environmental monitoring. Aust. J. Mar. Freshwat. Res. 42: 555-567.

Faith, D.P. 1990. Benthic macroinvertebrates in biological surveillance: Monte Carlo significance tests on functional groups' responses to environmental gradients. Environ. Monitor. Assess. 14: 247-264.

Faith, D.P., Humphrey, C.L. and Dostine, P.L. 1991. Statistical power and BACI designs in biological monitoring: Comparative evaluation of measures of community dissimilarity based on benthic macroinvertebrate communities in Rockhole Mine Creek, Northern Territory, Australia. Aust. J. Mar. Freshwat. Res. 42: 589-602.

Faith, D.P., Dostine, P.L. and Humphrey, C.L. 1995. Detection of mining impacts on aquatic macroinvertebrate communities: Results of a disturbance experiment and the design of a multivariate BACIP monitoring programme at Coronation Hill, Northern Territory. Aust. J. Ecol. 20: 167-180.

Ferraro, S.P., Cole, F.A., DeBen, W.A. and Swartz, R.C. 1989. Power-cost efficiency of eight macrobenthic sampling schemes in Puget Sound, Washington, U.S.A. Can. J. Fish. Aquat. Sci. 46: 2157-2165.

Fore, L.S., Karr, J.R. and Wisseman, R.W. 1996. Assessing macroinvertebrate responses to human activity: Evaluating alternative approaches. J. N. Amer. Benthol. Soc. 15(2): 212-231.

Frutiger, A. 1985. The production quotient, PQ -- a new approach for quality determination of slightly to moderately polluted running waters. Arch. Hydrobiol. 104: 513-526.

Goodman, D. 1975. The theory of diversity-stability relationships in ecology. Quart. Rev. Biol. 50: 237-266.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York, NY. 257 p.

Green, R.H. 1989. Power analysis and practical strategies for environmental monitoring. Environ. Res. 50: 195-205.

Green, R.H. and Montagna, P. 1996. Implications for monitoring: Study designs and interpretation of results. Can. J. Fish. Aquat. Sci. 53: 2629-2636.

Hart, D.D. and Robinson, C.T. 1990. Resource limitation in a stream community: Phosphorus enrichment effects of periphyton and grazers. Ecology 71: 1494-1502.

Hawkins, C.P. and Sedell, J.R. 1981. Longitudinal and seasonal changes in functional organization of macroinvertebrate communities in four Oregon streams. Ecology 62: 387-397.

Hawkins, C.P., Murphy, M.L. and Anderson, N.H. 1982. Effects of canopy, substrate composition, and gradients on structure of macroinvertebrate communities in Cascade Range streams of Oregon. Ecology 62: 387-397.

Heckman, C.W., Kamieth, H. and Stöhr, M. 1990. The usefulness of various numerical methods for assessing the specific effects of pollution on aquatic biota. Int. Revue ges. Hydrobiol. 75: 353-377.

Hellawell, J.M. 1977. Change in natural and managed ecosystems: Detection, measurement and assessment. Proc. Roy. Soc. Lond. B. 197: 31-57.

Hershey, A.E., Hiltner, A.L., Hullar, M.A.J., Miller, M.C., Vestal, J.R., Lock, M.A., Rundle, S. and Peterson, B.J. 1988. Nutrient influence on a stream grazer: *Orthocladius* microcommunities respond to nutrient input. Ecology 69: 1383-1392.

Hilsenhoff, W.L. 1977. Use of arthropods to evaluate water quality of streams. Tech. Bull. 100, Dept. of Natural Resources, Madison, WI. 15 p.

Hilsenhoff, W.L. 1987. An improved biotic index of organic stream pollution. Great Lakes Entomol. 20: 31-39.

Hruby, T. 1987. Using similarity measures in benthic impact assessments. Environ. Monitor. Assess. 8: 163-180.

Humphrey, C.L., Faith, D.P. and Dostine, P.L. 1995. Baseline requirements for assessment of mining impact using biological monitoring. Aust. J. Ecol. 20: 150-166.

Hurlbert, S.H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52: 577-586.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54: 187-211.

Hyland, J., Hardin, D., Steinhauer, M., Coats, D., Green, R. and Neff, J. 1994. Environmental impact of offshore oil development on the outer continental shelf and slope off Point Arguello, California. Mar. Environ. Res. 37: 195-229.

Johannsson, O.E. and Minns, C.K. 1987. Examination of association indices and formulation of a composite seasonal dissimilarity index. Hydrobiologia 150: 109-121.

Johnson, R.K., Wiederholm, T. and Rosenberg, D.M. 1993. Freshwater biomonitoring using individual organisms, populations, and species assemblages of benthic macroinvertebrates. *In*: Rosenberg, D.M. and Resh, V.H. (editors) 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, NY. p. 40-158.

Jones, J.R., Tracy, B.H., Sebaugh, J.L., Hazelwood, D.H. and Smart, M.M. 1981. Biotic index testing for ability to assess water quality of Missouri Ozark streams. Trans. Amer. Fish. Soc. 110: 627-637.

Keough, M.J. and Quinn, G.P. 1991. Causality and the choice of measurements for detecting human impacts in marine environments. Aust. J. Mar. Freshwat. Res. 42: 539-554.

Kerans, B.L. and Karr, J.R. 1994. A benthic index of biotic integrity for rivers of the Tennessee Valley. Ecol. Appl. 4: 768-785.

Klemm, D.J., Lewis, P.A., Fulk, F. and Lazorchak, J.M. 1990. Macroinvertebrate field and laboratory methods for evaluating the biological integrity of surface waters. Office of Research and Development, United States Environmental Protection Agency, Washington, D.C. EPA/600/4-90/030. 256 p.

Kiffney, P.M. and Clements, W.H. 1994. Effects of heavy metals on a macro-invertebrate assemblage from a Rocky Mountain stream in experimental microcosms. J. N. Amer. Benthol. Soc. 13: 511-523.

Kondratieff, P.F., Matthews, R.A. and Buikema, A.L. Jr. 1980. A stressed stream ecosystem: Macroinvertebrate community integrity and microbial trophic response. Hydrobiologia 111: 81-91.

Lamberti, G.A., Gregory, S.V., Ashkenas, L.R., Steinman, A.D. and McIntire, C.D. 1989. Productive capacity of periphyton as a determinant of plant-animal interactions in streams. Ecology 70: 1840-1856.

LaPoint, T.W., Melancon, S.M. and Morris, M.K. 1984. Relationships among observed metal concentrations, criteria, and benthic community structural responses in 15 streams. J. Water Pollut. Control Fed. 56: 1030-1038.

Leland, H.V., Fend, S.V., Dudley, T.L. and Carter, J.L. 1989. Effects of copper on species composition of benthic insects in a Sierra Nevada, California, stream. Freshwat. Biol. 21: 163-180.

Lenat, D.R. 1988. Water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. J. N. Amer. Benthol. Soc. 7: 222-233.

Lenat, D.R. 1993. A biotic index for the southwestern United States: Derivation and list of tolerance values, with criteria for assigning water-quality ratings. J. N. Amer. Benthol. Soc. 7: 222-233.

Loeb, S.L. and Spacie, A. (editors). 1994. Biological monitoring of aquatic systems. Lewis Pub., Boca Raton, FL. 381 p.

Lugthart, G.J. and Wallace, J.B. 1992. Effects of disturbance on benthic functional structure and production in mountain streams. J. N. Amer. Benthol. Soc. 11: 138-164.

MacArthur, R. 1955. Fluctuations of animal populations and a measure of community stability. Ecology 36: 533-536.

Maher, W.A. and Norris, R.H. 1990. Water quality assessment programs in Australia: Deciding what to measure, and how and where to use bioindicators. Environ. Monitor. Assess. 14: 115-130.

Manly, B.F.J. 1991. Randomization and Monte Carlo methods in biology. Chapman & Hall.

Marchant, R., Mitchell, P. and Norris, R. 1984. Distribution of benthic invertebrates along a disturbed section of the LaTrobe River, Victoria: An analysis bases on numerical classification. Aust. J. Mar. Freshwat. Res. 35: 355-374.

Margalef, R. 1968. Perspectives in ecological theory. University of Chicago Press, Chicago, IL. 111 p.

Mayack, D.T. and Waterhouse, J.S. 1983. The effects of low concentrations of particulates from paper mill effluent on the macroinvertebrate community of a fast-flowing stream. Hydrobiologia 107: 271-282.

Merritt, R.W. and Cummins, K.W. (editors). 1984. An introduction to the aquatic insects of North America. Second ed. Kendall/Hunt Pub. Co., Dubuque, Iowa. 722 p.

Metcalfe, J.L. 1989. Biological water quality assessment of running waters based on macroinvertebrate communities: History and present status in Europe. Environ. Pollut. 60: 101-139.

Minshall, G.W., Brock, J.T. and LaPoint, T.W. 1982. Characterization and dynamics of benthic organic matter and invertebrate functional feeding group relationships in the Upper Salmon River, Idaho (USA). Int. Revue ges. Hydrobiol. 67: 793-820.

Narf, R.P., Lange, E.L. and Wildman, R.C. 1984. Statistical procedures for applying Hilsenhoff's Biotic Index. J. Freshwat. Ecol. 2: 441-448.

Norris, R.H. 1995. Biological monitoring and the dilemma of data analysis. J. N. Amer. Benthol. Soc. 14: 440-460.

Norris, R.H. and Georges, A. 1993. Analysis and interpretation of benthic macroinvertebrate surveys. *In:* Rosenberg, D.M. and Resh, V.H. (editors) 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, NY. p. 234-286.

Novak, M.A. and Bode, R.W. 1992. Percent model affinity: A new measure of macroinvertebrate community composition. J. N. Amer. Benthol. Soc. 11: 80-85.

Odum, E.D. 1969. The strategy of ecosystem development. Science 164: 262-270.

Olive, J.H., Jackson, J.L., Bass, J., Holland, L. and Savisky, T. 1988. Benthic macroinvertebrates as indexes of water quality in the Upper Cuyahoga River. Ohio J. Science 88: 91-98.

Osborne, L.L., Davies, R.W. and Linton, K.J. 1980. Use of hierarchical diversity indices in lotic community analysis. J. Appl. Ecol. 17: 567-580.

Palmer, C.G., Maart, B., Palmer, A.R. and O'Keeffe, J.H. 1996. An assessment of macroinvertebrate functional groups as water quality indicators in the Buffalo River, eastern Cape Province, South Africa. Hydrobiologia 318: 153-164.

Peckarsky, B.L. and Cook, K.Z. 1981. Effect of Keystone Mine effluent on colonization of stream benthos. Environ. Entomol. 10: 864-871.

Perkins, J.L. 1983. Bioassay evaluation of diversity and community comparison indexes. J. Water Pollut. Control Fed. 55: 522-530.

Peterman, R.M. 1990a. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci. 47: 2-15.

Peterman, R.M. 1990b. The importance of reporting statistical power: The forest decline and acidic deposition example. Ecology 71: 2024-2027.

Peterson, C.H. 1993. Improvement of environmental impact analysis by application of principles derived from manipulative ecology: Lessons from coastal marine case histories. Aust. J. Ecol. 18: 21-52.

Peterson, C.H., Kennicutt, M.C. II, Green, R.H., Montagna, P., Harper, D.E. Jr., Powell, E.N. and Roscigno, P.F. 1996. Ecological consequences of environmental perturbations associated with offshore hydrocarbon production: A perspective on long-term exposures in the Gulf of Mexico. Can. J. Fish. Aquat. Sci. 53: 2637-2654.

Plafkin, J.L., Barbour, M.T., Porter, K.D., Gross, S.K. and Hughes, R.M. 1989. Rapid bioassessment protocols for use in streams and rivers. Benthic macroinvertebrates and fish. Office of Water Regulation and Standards, United States Environmental Protection Agency, Washington, D.C. EPA/440/4-89/001.

Pontasch, K.W. and Brusven, M.A. 1988. Diversity and community comparison indices: Assessing macroinvertebrate recovery following a gasoline spill. Water Res. 22: 619-626.

Pontasch, K.W., Smith, E.P. and Cairns, J. Jr. 1989. Diversity indices, community comparison indices and canonical discriminant analysis: Interpreting the results of multispecies toxicity tests. Water Res. 23: 1229-1238.

Poulton, B.C., Monda, D.P., Woodward, D.F., Wildhaber, M.L. and Brumbaugh, W.G. 1995. Relations between benthic community structure and metals concentrations in aquatic macroinvertebrates: Clark Fork Montana. J. Freshwat. Ecol. 10: 277-294.

Quinn, J.M. and Hickey, C.W. 1990. Characterisation and classification of benthic invertebrate communities in 88 New Zealand rivers in relation to environmental factors. N.Z. J. Mar. Freshwat. Res. 24: 387-409.

Rabeni, C.F., Davies, S.P. and Gibbs, E. 1985. Benthic invertebrate response to pollution abatement: Structural changes and functional implications. Water Res. Bull. 21: 489-498.

Rasmussen, K. and Lindegaard, C. 1988. Effects of iron compounds on macroinvertebrate communities in a Danish lowland river system. Water Res. 22: 1101-1108.

Reice, S.R. and Wohlenberg, M. 1993. Monitoring freshwater benthic invertebrates and benthic processes: Measures for assessment of ecosystem health. *In:* Rosenberg, D.M. and Resh, V.H. (editors) 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, NY. p. 287-305.

Resh, V.H. and Jackson, J.K. 1993. Rapid assessment approaches to biomonitoring using benthic macroinvertebrates. *In:* Rosenberg, D.M. and Resh, V.H. (editors) 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, NY. p. 195-233.

Resh, W.H., Norris, R.H. and Barbour, M.T. 1995. Design and implementation of rapid assessment approaches for water resource monitoring using benthic macroinvertebrates. Aust. J. Ecol. 20: 108-121.

Richardson, J.S. and Mackay, R.J. 1991. Lake outlets and the distributions of filter-feeders: An assessment of hypotheses. Oikos 62: 370-380.

Roline, R.A. 1988. The effects of heavy metals pollution of the upper Arkansas River on the distribution of aquatic macroinvertebrates. Hydrobiologia 160: 3-8.

Rosenberg, D.M. and Resh, V.H. (editors). 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, NY. 488 p.

Schindler, D.W. 1987. Detecting ecosystem responses to anthropogenic stress. Can. J. Fish. Aquat. Sci. 44 (Suppl. 1): 6-25.

Shaeffer, D.J. and Perry, J.A. 1986. Gradients in the distribution of riverine benthos. Freshwat. Biol. 16: 745-757.

Shannon, C.E. and Weaver, W. 1949. The mathematical theory of communication. University of Illinois Press, Urbana, IL. p. 19-27, 82-83, 104-107.

Simpson, E.H. 1949. Measurement of diversity. Nature 163(4148): 688.

Smith, B. and Wilson, J.B. 1996. A consumer's guide to evenness indices. Oikos 76: 70-82.

Smith, E.P., Orvos, D.R. and Cairns, J. Jr. 1993. Impact assessment using the before-after-control-impact (BACI) model: Concerns and comments. Can. J. Fish. Aquat. Sci. 50: 627-637.

Smith, E.P., Pontasch, K.W. and Cairns, J. Jr. 1990. Community similarity and the analysis of multispecies environmental data: A unified statistical approach. Water Res. 24: 507-514.

Sokal, R.R. and Rohlf, F.J. 1995. Biometry: The principles and practice of statistics in biological research. 3rd ed. W.H. Freeman & Co.

Stark, J.D. 1993. Performance of the Macroinvertebrate Community Index: Effects of sampling method, sample replication, water depth, current velocity, and substratum on index values. N.Z. J. Mar. Freshwat. Res. 27: 463-478.

Stewart-Oaten, A., Bence, J.R. and Osenberg, C.W. 1992. Assessing effects of unreplicated perturbations: No simple solutions. Ecology 73: 1396-1404.

Stewart-Oaten, A., Murdoch, W.W. and Parker, K.R. 1986. Environmental impact assessment: "Pseudoreplication" in time? Ecology 67: 929-940.

Suter, G.W. 1993. A critique of ecosystem health concepts and indexes. Environ. Toxicol. Chem. 12: 1533-1539.

Tabachnick, B.G. and Fidell, L.S. 1983. Using multivariate statistics. Harper & Row.

Taylor, B.R. 1997. Optimization of field and laboratory methods for benthic invertebrate monitoring. Canada Centre for Mineral and Energy Technology (CANMET), Natural Resources Canada, Ottawa, Ontario. 100 p.

Taylor, B.R. and Roff, J.C. 1986. Long-term effects of highway construction on the ecology of a southern Ontario stream. Environ. Pollut. (Series A) 40: 317-344.

Tuchman, N.C. and King, R.H. 1993. Changes in mechanisms of summer detritus processing between wooded and agricultural sites in a Michigan headwater stream. Hydrobiologia 268: 115-127.

Underwood, A.J. 1991. Beyond BACI: Experimental designs for detecting human environmental impacts on temporal variations in natural populations. Aust. J. Mar. Freshwat. Res. 42: 569-587.

Underwood, A.J. 1992. Beyond BACI: The detection of environmental impacts on populations in the real, but variable world. J. Exp. Mar. Biol. Ecol. 161: 145-178.

Underwood, A.J. 1994. On beyond BACI: Sampling designs that might reliably detect environmental disturbances. Ecol. Appl. 4: 3-15.

Van Hassel, J.H., Cherry, D.S., Hendricks, J.C. and Specht, W.L. 1988. Discrimination of factors influencing biota of a stream receiving multiple-source perturbations. Environ. Pollut. 55: 271-287.

Wallace, J.B. and Gurtz, M.E. 1986. Response of *Baetis* mayflies (Ephemeroptera) to catchment logging. Amer. Midl. Nat. 115: 24-41.

Wallace, J.B. and Webster, J.R. 1996. The role of macroinvertebrates in stream ecosystem structure. Ann. Rev. Entomol. 41: 115-140.

Wallace, J.B., Cuffney, T.F., Webster, J.R., Lugthart, G.J., Chung, K. and Goldowitz, B.S. 1991. Export of fine organic particles from headwater streams: Effects of season, extreme discharges, and invertebrate manipulation. Limnol. Oceanogr. 36: 670-682.

Wallace, J.B., Grubaugh, J.W. and Whiles, M.R. 1996. Biotic indices and stream ecosystem processes: Results from an experimental study. Ecol. Appl. 6: 140-151.

Warwick, R.M. 1993. Environmental studies on marine communities: Pragmatical considerations. Aust. J. Ecol. 18: 63-80.

Warwick, R.M. and Clarke, K.R. 1991. A comparison of some methods for analysing changes in benthic community structure. J. Mar. Biol. Assoc. U.K. 71: 225-244.

Warwick, R.M. and Clarke, K.R. 1993. Comparing the severity of disturbance: A meta-analysis of marine macrobenthic community data. Mar. Ecol. Prog. Series 92: 221-231.

Washington, H.G. 1984. Diversity, biotic and similarity indices: A review with special relevance to aquatic ecosystems. Water Res. 18: 653-694.

Waterhouse, J.C. and Farrell, M.P. 1985. Identifying pollution related changes in chironomid communities as a function of taxonomic rank. Can. J. Fish. Aquat. Sci. 42: 406-413.

Wilhm, J.L. and Dorris, T.C. 1968. Biological parameters for water quality criteria. BioScience 18: 477-481.

Winner, R.W., Boesel, M.W. and Farrell, M.P. 1980. Insect community structure as an index of heavy-metal pollution in lotic ecosystems. Can. J. Fish. Aquat. Sci. 37: 647-655.

Woodiwiss, F.S. 1964. The biological system of stream classification used by the Trent River Board. Chemistry and Industry 83: 443-447.

Zar, J.H. 1996. Biostatistical analysis. 3rd ed. Prentice Hall, Englewood Cliffs, N.J.